(12)

ADVANCED TELEPROCESSING SYSTEMS
DEFENSE ADVANCED RESEARCH PROJECTS AGENCY

SEMI-ANNUAL TECHNICAL REPORT

SEPTEMBER 30, 1983

AD A137944

Principal Investigator: Leonard Kleinrock

DTIC
ELECTE
FEB 15 1984
S D
B

Computer Science Department
School of Engineering and Applied Science
University of California
Los Angeles

DTIC FILE COPY

DISTRIBUTION STATEMENT A
Approved for public release;
Distribution Unlimited

84 02 14 030

UNIVERSITY OF CALIFORNIA, LOS ANGELES

SEMI-ANNUAL TECHNICAL REPORTS

Sponsored by the

Defense Advanced Research Projects Agency

Contract Numbers: DAHC-15-C-0368
MDA 903-77-C-0272
MDA 903-82-C-0064

DARPA Order No. 2496

COMPUTER NETWORK RESEARCH

DATES	DDC ACCESSION NUMBER
August 1969 to February 1970:	AD 705 149
August 1970:	AD 711 342
June 1971:	AD 727 989
December 1971:	AD 739 705
June 1972:	AD 746 509
December 1972:	AD 756 708
June 1973:	AD 769 706
December 1973:	AD A004167
June 1974:	AD A008422
December 1974:	AD A016823
June 1975:	AD A020671
December 1975:	AD A025914
June 1976:	AD A034171 (Final)

ADVANCED TELEPROCESSING SYSTEMS

June 1976 to December 1976:	AD A039018
June 1977:	AD A047496
June 1978:	AD A077404
September 1979:	AD A081938
March 1980:	AD A088839
September 1981:	AD A133525 (Final)
September 1982:	to be assigned
March 1983:	to be assigned

ADVANCED TELEPROCESSING SYSTEMS

Semi-Annual Technical Report

September 30, 1983

Contract Number: MDA 903-82-C-0064
DARPA Order Number: 2496
Contract Period: October 1, 1981 to September 30, 1983
Report Period: ~~March 31~~ APRIL 1, 1983 to September 30, 1983

Principal Investigator: Leonard Kleinrock

Co-Principal Investigator: Mario Gerla

(213) 825-2543

DTIC
ELECTE
FEB 15 1984
S D
B

Computer Science Department
School of Engineering and Applied Science
University of California, Los Angeles

Sponsored by

DEFENSE ADVANCED RESEARCH PROJECTS AGENCY

DISTRIBUTION STATEMENT A
Approved for public release
Distribution Unlimited

The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the Defense Advanced Research Projects Agency or the United States Government.

SECURITY CLASSIFICATION OF THIS PAGE *(When Data Entered)*

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER	2. GOVT ACCESSION NO. AD-A137 444	3. RECIPIENT'S CATALOG NUMBER
4. TITLE *(and Subtitle)* ADVANCED TELEPROCESSING SYSTEMS: SEMI-ANNUAL TECHNICAL REPORT		5. TYPE OF REPORT & PERIOD COVERED Semi-Annual Technical 1 APR 83 through 30 SEPT 83 6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Leonard Kleinrock		8. CONTRACT OR GRANT NUMBER(s) MDA 903-82-C-0064
9. PERFORMING ORGANIZATION NAME AND ADDRESS School of Engineering and Applied Science University of California, Los Angeles Los Angeles, CA 90024		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS DARPA Order No. 2496
11. CONTROLLING OFFICE NAME AND ADDRESS Defense Advanced Research Projects Agency 1400 Wilson Boulevard Arlington, VA 22209		12. REPORT DATE September 30, 1983 13. NUMBER OF PAGES 237
14. MONITORING AGENCY NAME & ADDRESS*(if different from Controlling Office)*		15. SECURITY CLASS. *(of this report)* 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT *(of this Report)*

Approved for Public Release; Distribution Unlimited

17. DISTRIBUTION STATEMENT *(of the abstract entered in Block 20, if different from Report)*

18. SUPPLEMENTARY NOTES

19. KEY WORDS *(Continue on reverse side if necessary and identify by block number)*

Computer network power, Computer communication networks, Throughput-delay performance measures, Terrestrial wire network, Network operating point, Flow control, Optimization, Packet Switching.

20. ABSTRACT *(Continue on reverse side if necessary and identify by block number)*

This Semi-Annual Technical Report covers research carried out by the Advanced Teleprocessing Systems Group at UCLA under DARPA Contract No. MDA 903-82-C-0064 covering the period from April 1, 1983 to September 30, 1983.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

S/N 0102-LF-014-6601

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE *(When Data Entered)*

(20)

This contract has three primary designated research areas: packet radio systems, resource sharing and allocation, and distributed processing and control.

This report contains the abstracts of the publications which summarize our research results in those areas during this semi-annual period, followed by the main body of the report which consists of the Ph.D. dissertation by H. Richard Gail, "On the Optimization of Computer Network Power," conducted under the supervision of Professor Leonard Kleinrock (Principal Investigator for this contract). It addresses the tradeoff between throughput and delay involving the selection of a suitable operating point for a computer network. This tradeoff is studied through the maximization of various throughput-delay performance measures, all known as power. The models analyzed for the most part are those for a terrestrial wire network. Power is first analyzed for simple computer networks. Although these networks are topologically simple, they also yield single-variable optimization problems which are mathematically simple. The critical system parameter turns out to be the average number in system at maximum power since it is a parameter which is easily implemented in networks which use window flow control and also because it exhibits important invariances. A network which is topologically simple but which no longer has a simple problem formulation is also studied. It is found that some of the nice results obtained for the simple formulations do not hold for this more complicated multi-variable problem. In particular, issues involving fairness of operating points are explored. The power problem is extended in several ways. First the problem formulation itself is altered (the constraints and/or decision variables); second the objective function (the power function) is changed. With these extensions, the multi-variable problem is often manageable. A generalized power performance measure, which enables the analyst to vary the importance of throughput relative to delay, for the above problem formulations is also studied. Finally, the analysis of power is extended to networks with blocking. Tradeoffs among throughput, delay and the blocking probability are examined.

DTIC

Accession For	
NTIS GRA&I	☑
DTIC TAB	☐
Unannounced	☐
Justification	
PER CALL JC	
By	
Distribution/	
Availability Codes	
Dist	Avail and/or Special
A-1	

ADVANCED TELEPROCESSING SYSTEMS

Defense Advanced Research Projects Agency
Semi-Annual Technical Report

September 30, 1983

INTRODUCTION

This Semi-Annual Technical Report covers research carried out by the Advanced Teleprocessing Systems Group at UCLA under DARPA Contract No. MDA 903-82-C-0064 covering the period from April 1, 1983 through September 30, 1983. Under this contract we have three designated tasks as follows:

TASK I. PACKET RADIO SYSTEMS

The extension of our analytic and design techniques to modern multi-hop packet radio networks will be studied. The applications and extensions include access methods, large network control and management, queueing network models, approximation methods, capture phenomena, conflict-free algorithms, reliability, routing procedures, topological studies, TDMA in a multi-hop environment and multiplexing methods.

TASK II. RESOURCE SHARING AND ALLOCATION

Extended concepts of "power" in networks will be studied. The extensions include more complex topologies and configurations, extended queueing disciplines, general distributions, other definitions of power, effects of varying the traffic matrix and fairness. The problems of large scale internetting with respect to resource allocation and sharing will also be studied further.

TASK III. DISTRIBUTED PROCESSING AND CONTROL

Overall principles of distributed processing and distributed control will be studied. The issues of sequencing in data base updates, distributed control and distributed processing (involving the calculation of concurrency of processing) are the subjects of concern here.

A major contribution of our research during this reporting period is contained in Reference 7 listed below, namely, "On the Optimization of Computer Network Power," by H. Richard Gail. This dissertation was supervised by Professor Leonard Kleinrock (Principal Investigator for this research). It addresses the tradeoff between throughput and delay involving the selection of a suitable operating point for a computer network. This tradeoff is studied through the maximization of various throughput-delay performance measures, all known as *power*. The models

analyzed for the most part are those for a terrestrial wire network. Power is first analyzed for simple computer networks. Although these networks are topologically simple, they also yield single-variable optimization problems which are mathematically simple. The critical system parameter turns out to be the average number in system at maximum power since it is a parameter which is easily implemented in networks which use window flow control and also because it exhibits important invariances. A network which is topologically simple but which no longer has a simple problem formulation is also studied. It is found that some of the nice results obtained for the simple formulations do not hold for this more complicated multi-variable problem. In particular, issues involving fairness of operating points are explored. The power problem is extended in several ways. First, the problem formulation itself is altered (the constraints and/or decision variables); second, the objective function (the power function) is changed. With these extensions, the multi-variable problem is often manageable. A generalized power performance measure, which enables the analyst to vary the importance of throughput relative to delay for the above problem formulations, is also studied. Finally, the analysis of power is extended to networks with blocking. Tradeoffs among throughput, delay and the blocking probability are examined. The entire dissertation is reproduced as the main body of this report. The following list of research publications summarizes the results of this semi-annual period and the abstract of each paper is given along with the reference itself.

RESEARCH PUBLICATIONS

1. **Nelson, R. and L. Kleinrock, "Maximum Probability of Successful Transmission in a Random Planar Packet Radio Network,"** *Infocom '83 Proceedings,* **April 18-21, 1983, San Diego, California.**

Suppose a packet radio network has nodes which are randomly distributed over the infinite plane according to a Poisson point process such that nodes have an average of N nodes within its transmission range. In this paper we show that, over all protocols, the maximum probability of a successful transmission over any period of time is upper bounded by $.9278/N$ for suitably large N. We compare this performance to that obtained using slotted-ALOHA and CSMA and show, for realistic networks, that these protocols at best achieve respectively about 36% and 49% of our bound.

2. **Kleinrock, L., "Packet Switching Principles,"** *Proceedings of the L. M. Ericsson Award Ceremony,* **Stockholm, Sweden, May 5, 1982, also published as a special editorial in** *Journal of Telecommunication Networks,* **Spring 1983, pp. 1-5.**

This paper traces the development of packet switching over the last decade, explores its underlying principles and describes its impact on the modern world of data communications. The likely applications to advanced future systems is also discussed.

3. **Sadr, R., "UCLA Demodulation Engine," UCLA Computer Science Department Report No. CSD-830518, May 1983.**

This report describes the VLSI design and implementation of a Viterbi algorithm processor for simultaneous data demodulation and phase tracking of Minimum Shift Keying signal.

During the 1981-82 academic year, graduate students in the VLSI course (CS258A-C) at UCLA designed the implementation of this system as a one-year class project and, with support from DARPA (Defense Advanced Research Project Agency), fabricated this processor on a single chip, using 4-micron NMOS technology. The *UCLA Demodulation Engine* can be used as an inexpensive digital radio receiver in a variety of applications.

4. **Sadr, R., "Receiver Design and Analysis for Generalized Minimum Shift Keying Modulation Techniques," Ph.D. Dissertation, University of California, Los Angeles, June 1983.**

In this dissertation we consider the design, analysis and implementation of optimum demodulators for a class of coherent Minimum Shift Keying (MSK) signals. An open loop design for simultaneous data demodulation and phase tracking of the MSK signal using the Viterbi algorithm is considered.

Two variations of the MSK signal are studied. The MSK with overlay is a dual rate modulation technique which is equivalent to transmitting two symbols during certain periods of the MSK signal. The demodulator uses the Viterbi algorithm to estimate both the low and high rate data simultaneously from the data signal. The MSK with Pseudo Random (PN) sequence combats intentional or unintentional jamming in spread spectrum systems. We find a simplified receiver for MSK with PN sequence and also a demodulator which takes into consideration the effect of random phase perturbations.

The performance of these demodulators is evaluated using new transfer function bounds for periodically time varying finite state dynamical systems and can also be applied for time varying codes and modulations.

The VLSI system architecture of the Viterbi algorithm processor for simultaneous data demodulation and phase tracking of the MSK signal is formulated. The UCLA course on VLSI during the 1981-82 academic year undertook the implementation of the UCLA Demodulation Engine as a one year class project. This chip is fabricated on a single die, using 4-micron NMOS technology.

5. **Silvester, J. A. and L. Kleinrock, "On the Capacity of Multi-Hop Slotted ALOHA Networks with Regular Structure,"** *IEEE Transactions on Communications,* **Vol. COM-31, August 1983, pp. 974-982.**

> In this paper we investigate the capacity of networks with a regular structure operating under the slotted ALOHA access protocol. We first consider circular (loop) and linear (bus) networks and then proceed to two-dimensional networks. For one-dimensional networks we find that the capacity is basically independent of the network average degree and is almost constant with respect to network size. For two-dimensional networks we find that the capacity grows in proportion to the square root of the number of nodes in the network, provided that the average degree is kept small. Furthermore, we find that reducing the average degree (with certain connectivity restrictions) allows a higher throughput to be achieved. We also investigate some of the peculiarities of routing in these networks.

6. **Silvester, J. A. and L. Kleinrock, "On the Capacity of Single-Hop Slotted ALOHA Networks for Various Traffic Matrices and Transmission Strategies,"** *IEEE Transactions on Communications,* **Vol. COM-31, August 1983, pp. 983-991.**

> In this paper we formulate a general model of the capacity of single-hop slotted ALOHA networks. We find that the capacity can be expressed as a function of the nodal degree (i.e., number of nodes within range of a transmitter). We then evaluate this model for various traffic matrices. In order to satisfy the requirements of a given traffic matrix, the transmission power is selected accordingly and this determines the degree of the nodes and, hence, the network performance. Finally, we compare our results to simulation studies.

7. **Gail, H. Richard, "On the Optimization of Computer Network Power," Ph.D. Dissertation, Computer Science Department, University of California, Los Angeles, September 1983, Report No. CSD-830922.**

September 1983

ON THE OPTIMIZATION OF COMPUTER NETWORK POWER

by

H. Richard Gail

This research, conducted under the chairmanship of Professor Leonard Kleinrock, was sponsored by the Defense Advanced Research Projects Agency, Department of Defense.

Computer Science Department
School of Engineering and Applied Science
University of California
Los Angeles

ABSTRACT

We study the tradeoff between throughput and delay involving the selection of a suitable operating point for a computer network. We choose to analyze this tradeoff through the maximization of various throughput-delay performance measures, all known as *power*. The models we analyze for the most part are those for a terrestrial wire network.

We begin by analyzing power for simple computer networks. Although these networks are topologically simple, they also yield single-variable optimization problems which are mathematically simple. We concentrate on a critical system parameter, the average number in system at maximum power, since it is a parameter which is easily implemented in networks which use window flow control, and also because it exhibits important invariances. We next analyze a network which is topologically simple but which no longer has a simple problem formulation. We find that some of the nice results obtained for the simple formulations do not hold for this more complicated multi-variable problem. In particular, issues involving fairness of operating points are explored.

We then extend the power problem in several ways. First the problem formulation itself is altered (the constraints and/or decision variables); second the objective function (the power function) is changed. We find with these extensions, that the multi-variable problem is often manageable. We also study a generalized power performance measure, which enables the analyst to vary the importance of throughput relative to delay, for the above problem formulations. Finally the analysis of power is extended to networks with blocking. Tradeoffs among throughput, delay and the blocking probability are examined.

Table of Contents

List of Figures

page

CHAPTER 1
Introduction

How does one select an "appropriate" operating point for a computer network? This seemingly simple question does not have a straightforward answer. The object of this research is to contribute toward answering the above question.

1.1 Computer Network Models

Before addressing this general question, we must describe the modeling environment of computer networks. It has long been recognized that sharing of the various network resources (channels, nodes, software, etc.) is essential since data traffic is known to be "bursty" in nature. In conventional land-based wire networks, customers (messages, packets) enter the system and are served (transmitted) by the interconnecting message channels. In broadcast networks (satellite, packet radio, broadcast cable), collisions of messages may occur since it is no longer possible to schedule (queue) customer requests. In this case, the system resources (the communication channels) may be used very inefficiently. To analyze both types of networks, various performance measures have been proposed. These include throughput, response time (delay), backlog, efficiency and system capacity. Other measures such as line cost, buffer size, reliability, blocking and fairness have also been studied.

The first modeling and analysis of a computer network was that of Kleinrock [Klei64]. Using various assumptions (such as external Poisson arrivals, infinite nodal storage, exponential message length and the well-known independence assumption) he was able to model a conventional wire network as a queueing network of the type studied by Jackson [Jack63]. In such Jacksonian networks, the analysis of certain network measures may be reduced to the study of the individual nodes. The system probabilities obey a product-form solution with respect to the nodal probabilities. Thus Kleinrock could find expressions for average number in system and also for average delay. He extended his studies to design issues, and for design variables he included the topology of the network, the capacities of the links, and the flows on the channels. The networks we will analyze in this dissertation will (for the most part) be of the above type, and thus the models used will be based on that of Kleinrock.

1.1.1 Notation

We now present the notation which we will use in this dissertation. It follows that employed by Kleinrock [Klei75, Klei76] in modeling conventional wire terrestrial computer networks. We assume the network has N nodes (switching computers) and M channels, and that the topology is arbitrary. The capacity of the ith channel is assumed to be C_i bits per second. The mean length of the messages which traverse the network is $\overline{b}$ bits per message (the message length distribution is sometimes assumed to be exponential). Each source-destination pair (j,k) of nodes of the network represents a potential user (or users) who wishes to send messages from (a HOST connected to) node j to (a HOST connected to) node k. We assume that the traffic originating from this user occurs with a rate of γ_{jk} messages per second. We usually consider these arrival processes to be *Poisson*, although that restriction will be relaxed on occasion. The total external arrival rate of messages to the system is simply

$$\gamma = \sum_{j=1}^{N} \sum_{k=1}^{N} \gamma_{jk} \tag{1.1}$$

In the networks considered here we usually assume messages are not blocked or lost (e.g. infinite buffer capacity, no noise on the lines and no possibility of collision as in packet radio networks or local networks of the ETHERNET type). Thus throughput is synonymous with the traffic applied to the system, and so it is equal to γ for these networks (however, the case of blocking will be considered in chapter 7 below).

The arrival rate of messages which are transmitted over channel i is assumed to be λ_i messages per second. Therefore, the total traffic within the network is

$$\lambda = \sum_{i=1}^{M} \lambda_i \tag{1.2}$$

We next set T to be the average total time spent in the network by a message. Thus T is the sum of W, the average total waiting time (on queues) spent in the network by a message, and $\overline{x}$, the average total service time of a message. (That is, $\overline{x}$ is the total time a message spends in transmission on all channels in its journey through the net.) We have

$$T = W + \overline{x} \tag{1.3}$$

Finally we set $\overline{N}$ to be the average total number of messages in the network.

These various network parameters are composed of corresponding quantities for each of the message channels in the network. We let $\overline{N}_i$, T_i, W_i, $\overline{x}_i$ be the values for channel i. Clearly

$$\overline{N} = \sum_{i=1}^{M} \overline{N}_i \tag{1.4}$$

The service time for a message on the ith channel, $\bar{x}_i$, may be expressed as the average length of a message, $\bar{b}$, in bits, divided by the capacity of the ith channel, C_i, in bits per second. That is, we have

$$\bar{x}_i = \frac{\bar{b}}{C_i} \tag{1.5}$$

Thus the variation in service time at the ith channel occurs due to the variation in message length. We also define the utilization (efficiency) for the ith channel as

$$\rho_i = \lambda_i \bar{x}_i \tag{1.6}$$

Using these quantities for the individual message channels, Kleinrock established the key delay throughput relationship

$$T = \sum_{i=1}^{M} \frac{\lambda_i}{\gamma} T_i \tag{1.7}$$

We also will use Little's result [Litt61] which relates several of the above quantities as

$$\bar{N} = \gamma T \tag{1.8}$$

Note that equation (1.7) is a consequence of equation (1.4) and Little's result. Summarizing these definitions we have the following list:

M	number of channels
N	number of nodes
C_i	capacity of the ith channel
$\bar{b}$	mean message length (bits)
γ_{jk}	traffic from node j to node k
γ	total external traffic (throughput)
λ_i	traffic on channel i
λ	total network traffic
T	total mean time in system
T_i	mean time at channel i
W	total mean waiting time (on queues)
W_i	mean waiting time at channel i
$\bar{x}$	total mean service time
$\bar{x}_i$	mean service time at channel i
$\bar{N}$	total mean number in system
$\bar{N}_i$	mean number at channel i
ρ_i	utilization (efficiency) for channel i

1.1.2 Delay as a Performance Measure

Kleinrock used equation (1.7) as the overall performance measure for a computer network. Further, he posed several optimization problems, three of which we now review since we intend to consider similar problems using a new performance measure called power throughout this dissertation. In all three problems, the topology of the network and the traffic matrix $\{\gamma_{jk}\}$ are assumed to be given.

(i) In the *capacity assignment problem (CA)* the channel flows $\{\lambda_i\}$ are given, and one seeks to select capacities which minimize total system delay T under a given capacity cost constraint. Using the method of Lagrange multipliers [Mang69], Kleinrock solved this problem assuming linear costs (e.g. channel cost proportional to capacity). The result is the so-called square root capacity assignment [Klei64]. This problem has also been analyzed assuming other cost functions [Klei76].

(ii) For the *flow assignment problem (FA)* the capacities are now given and the channel flows must be determined which minimize T. The problem is a convex cost multicommodity flow problem with the interesting fact that upper bounds on the flows are incorporated (implicitly) into the objective function as penalties. Convex cost network flow problems also occur in other diverse areas (urban transportation systems, pipe network systems and production-distribution problems), and there is a considerable literature on the attempts to solve them [Kenn80]. We mention two methods tailored to the communications design problem, namely the *flow deviation algorithm* [Frat73, Klei76] and the *extremal flows method* [Cant74]. The flow deviation method is a variant of the Frank-Wolfe algorithm, while the extremal flows method is an application of Dantzig-Wolfe decomposition. The shortest route subproblems of flow deviation can be solved very quickly with programming techniques that emphasize the use and manipulation of certain data structures [Kenn80].

(iii) As a final example, the above two problems have been extended in a natural way to obtain the *capacity and flow assignment problem (CFA)*. Here the decision variables include both capacities and flows. A suboptimal algorithm has been studied [Frat73, Klei76] which, in the language of Geoffrion [Geof70], makes use of the technique called *projection*. The projected problem turns out to be a non-convex programming problem on a convex polyhedron.

Clearly, delay must be considered when selecting a network operating point.

1.1.3 Throughput as a Performance Measure

We have given three examples of network design problems for wire networks that have varying degrees of difficulty. All use delay as the principal measure of interest. Another key measure of performance is system throughput which we discuss in this section.

Let us focus on two issues of importance in the analysis of the throughput of real systems, those of *routing* and *flow control*. We have seen an example of a (static) routing problem, namely the flow assignment problem; the task was to route traffic (i.e. find the channel flows) in such a way so as to minimize total system delay. In real systems the bursty nature of the traffic may require dynamic routing algorithms to enable the network to adjust to the changing traffic, as well as dynamic flow control schemes to provide high throughput.

In discussing the routing problem we assumed a certain amount of traffic had to be sent through the network in an efficient manner. Of course it is possible that too many messages may be in the network thus straining the capacity of various system resources. Congestion, throughput degradation and even deadlock may develop. The system may not be able to function under heavy load and may grind to a halt. One answer to this problem is to throttle traffic entering the network (or perhaps at intermediate nodes). A set of protocols to achieve this is called a flow control scheme, some of which have been (heuristically) developed. Both static and dynamic flow control schemes have been studied [Klei80, Kerm80]. A recent survey of the various levels of flow control which may be needed in a network are discussed in [Gerl80]. The book [Gran79] edited by Grangé and Gien contains the proceedings of a symposium devoted entirely to the flow control problem.

How are these issues relevant to the choice of an operating point? Once we somehow determine the characteristics such a point should have (using some performance criterion say), flow control is one mechanism that might be used in attempting to achieve this goal. For example, if we claim that for a certain network the operating point occurs when the rate of messages into the system is λ^* (msg/sec), or perhaps it occurs when the average number of messages in the system is $\overline{N}^*$, then an appropriate flow control strategy may be considered in order to operate the network at the right point. The design of a good flow control strategy (using permits, windows, etc.) will depend on the physical characteristics of the network in question. Figure 2 of [Klei78a] gives a nice graphical interpretation of the interaction of throughput with a flow control scheme.

Throughput is also an important performance measure for broadcast networks, often considered in the literature to be more important than delay. In the single-hop broadcast environment (satellite, etc.), protocol capacity (that is, the amount of throughput that can be achieved if we pump the system until the point of infinite delay) seems to be the current measure in favor [Moll82]. Perhaps this is partially due to the complexity of the delay analysis for such networks. As a result, the emphasis in this area has been to invent an access scheme and see how much throughput it gives if you run it to capacity. These results are quite interesting, but they fail to answer the question of where to operate a single-hop broadcast system.

In the multi-hop broadcast packet radio network, other performance measures have been examined [Nels82]. In such systems individual nodes have the ability to hear a subset of the set of all nodes, but as in single-hop systems collisions can occur. This has some of the graphical flavor of the wire network problem combined with some aspects of single-hop. Two measures of interest are the expected number of successful transmissions (perhaps normalized) and the expected progress (in hops) per transmission attempt. In some sense the first is a measure of throughput while the second is a measure of delay (if the average length a message must travel is $\bar{n}$ hops, for example, we may obtain an indication of delay using the expected progress criterion).

1.1.4 Throughput-Delay Tradeoffs

We have reviewed various physical networks and the techniques used to model and analyze them. Although several performance measures have been discussed in the literature, no one single measure gives an adequate answer to the main question posed at the beginning of this chapter. Two important measures, throughput and delay, have emerged which, in some sense, are conflicting. If the input rate of messages to a network is decreased in order to decrease the system delay, the throughput will also decrease. Conversely, if the rate of messages to a network is increased to obtain higher throughput, the delay will increase. Thus one might suggest that an appropriate operating point may incorporate some type of tradeoff between the competing measures of throughput and delay. Several such tradeoff functions have in fact been introduced.

One such function, called *power,* has been found to be quite useful, and it is the behavior of this measure (and its generalizations) which is addressed in this dissertation. In its simplest form, power is the ratio of throughput and delay (i.e., $P = \gamma/T$). Thus P is an increasing function of the (good) performance measure throughput and a decreasing function of the (bad) performance measure delay.

1.2 Summary of Results

We now give a brief outline of this dissertation and summarize its major results. In this first chapter, we have discussed several previous models of computer networks and the accompanying analyses. Such analytical work is necessary before we can attempt to select a network operating point. The list of models is by no means exhaustive, but illustrates the type of performance measures used in the study of such networks and represents candidates for the extension of this research.

In chapter 2 a recently introduced concept called *power* is discussed. Power, P, is a measure which combines the two most common performance measures (throughput γ and delay T) into a single performance function as $P = \gamma/T$. Several simple networks are analyzed from the power point of view and intuitive rules of thumb are shown to be valid for these

networks. The networks which are studied are not only topologically simple, but they yield problems which are simple from an optimization point of view. Of major interest is the parameter $\overline{N}^*$, the average number in system at the optimal power point. Not only is it implementable in a window flow control scheme, but we also find that it is invariant under scaling of channel capacities in some cases, and invariant with respect to the service time distribution in other cases. We first study two types of *series* networks with Poisson input, the M/M/1 series network (exponential message length) and the M/D/1 series network (constant message length), and obtain results for arbitrary channel capacities. In the particular case of equal channel capacities, we show that $\overline{N}^*$ is identical for both networks, even though other system parameters (such as the optimal throughput γ^*) are quite different. We next consider a *parallel* network with Poisson input and arbitrary message length, the M/G/1 parallel network. After obtaining several results for this general parallel network, we consider two special cases: exponential message length (the M/M/1 parallel network); and equal loads on each channel. We show that the equation

$$\overline{N}^* = \sum_{i=1}^{M} \overline{N}_i^* = \sum_{i=1}^{M} (\overline{N}_i^*)^2 \tag{1.9}$$

characterizing $\overline{N}^*$, which was given by Bharath-Kumar [Bhar80] for the M/M/1 series network, also holds for the M/M/1 parallel network. When we have Poisson input to the network, we obtain various extensions of the "keep the pipe full" result of Kleinrock [Klei79] which states that $\overline{N}^* = 1$ for the M/G/1 queueing system. However, we then show that for any positive ϵ, there is a G/M/1 system with $0 < \overline{N}^* < \epsilon$, which is certainly different from the M/G/1 result.

In chapter 3 a topologically simple network (a parallel net) is analyzed, but the optimization problem which results is no longer mathematically simple. Although the input traffic is Poisson and the message length is exponential as was the case for the M/M/1 parallel network studied in chapter 2, the traffic on the M channels is optimized individually, yielding a multivariable optimization problem instead of the simple single-variable problems of chapter 2. We obtain an analytical solution for the case of two channels and give an optimization procedure for the general case. Equation (1.9), which characterized the M/M/1 networks of chapter 2, is again shown to hold. A notion of a fair operating point, due to Jaffe [Jaff81], is examined for this network (an operating point is fair if all users of the network receive positive throughput). For the two channel case, we show that the optimal power point is fair if and only if the ratio of the capacity of the fast channel to that of the slow channel is less than four, and fairness results are also presented for the general M channel case. Thus optimal power points may be unfair to certain users. We also give an example of a power function which is not concave. These undesirable properties motivate extension of the power analysis in two directions (which are discussed in the following chapter).

In chapter 4 the analysis of power is extended in two ways. In the first section of the chapter, we change the routing constraints and/or the traffic matrix of our power problem formulation. One particular formulation (known routing and relative traffic matrix) is solved completely for an M/M/1 network with arbitrary topology. Using this solution, we show that the equation which characterizes $\overline{N}^*$ for the M/M/1 series network and the M/M/1 parallel network (equation (1.9) above) holds for all the formulations introduced in this chapter. In the second section of chapter 4, the objective function of our optimization problem is changed. Several extensions of the definition of power which appeared previously in the literature are compared and contrasted. We show that one power function (first introduced by Kleinrock) has properties that are more intuitively pleasing than the other definitions of power. For a general network power problem using Kleinrock's definition, we show that one should "keep the pipe full" (i.e., $\overline{N}_i^* = 1$ for all i) if the constraints of the problem allow it. We also give an example of a solution which optimizes Kleinrock's power function and does "keep the pipe full", but which is unfair.

In chapter 5, we study a performance measure introduced by Kleinrock, called generalized power, which allows the analyst to vary the importance of throughput relative to delay. This new family of functions is then analyzed for the simple network problems introduced above. Extensions of results in chapter 2 and chapter 4 are obtained.

In chapter 6, the analysis of generalized power is extended to the parallel network studied in chapter 3. Two different families of generalized power functions are examined, and one family is shown to have optimal solution points with desirable properties such as fairness.

In chapter 7 the study of power is extended to multiple-server systems and to systems with blocking. Under the influence of a smoothing principle for large shared systems (which is nothing more than the law of large numbers), the asymptotic behavior of large systems is observed to be deterministic in nature. Various deterministic rules of thumb are derived. Both pure loss systems and systems with finite waiting room are also studied. The previous definitions of power are shown to be inadequate for use with these blocking systems, and a further extension of power due to Kleinrock which incorporates the blocking probability is examined. Kleinrock's new definition not only is an increasing function of the (good) measure throughput and a decreasing function of the (bad) measure delay as before, it is also a decreasing function of the (bad) measure blocking. This new power function is shown to peak at the intuitively correct point for the limiting case of large pure loss systems.

In chapter 8 we list further proposed topics which might be carried out as extensions of this study.

CHAPTER 2

An Analysis of Power for Simple Computer Network Configurations

In this chapter we introduce the performance measure of interest to this dissertation, namely, *power,* several definitions of which have appeared in the literature. We then analyze various simple computer network configurations with respect to the maximization of power. The networks are not only topologically simple, but (what is perhaps more important) they are *mathematically simple* from an optimization point of view. The optimization problems which result involve a single decision variable, and equations which characterize the optimal power point are easily derived for these simple networks. These equations are then utilized to study several simple computer network models.

First a series network with Poisson arrivals and exponential message length (which can be used to model a path through a network with large mixing of messages) is analyzed. Next a series network with Poisson arrivals and constant message length (e.g., a path in a virtual circuit network) is studied. The results for these two tandem models are contrasted. Then a parallel configuration with Poisson arrivals and general message length distribution is studied (e.g., a packet switch with numerous outgoing channels where the average number represents the buffer size or the window size). Several special cases of parallel networks for certain network parameters are considered.

For the simple networks studied, we show that, when the network is operated at the maximum power point, the analysis leads to a value for the average number in system which is invariant under scaling of line capacities in some cases and invariant under distribution of message length (service time) in other cases. This value of the average number in system is easily calculated for the simple networks we consider, and controlling it in an operational network may be implemented using a window flow control scheme [Gerl80].

Finally the Poisson arrival assumption is dropped, and a queueing system with general input and exponential service is analyzed. Many of the nice results which are true under Poisson arrivals are shown to no longer hold for systems which have a more general arrival process.

2.1 Power as a Performance Measure

Our interest is the tradeoff between throughput and delay involved in choosing a particular system operating point. As the input traffic (messages, packets) offered to a network increases, the mean system delay increases. One might also expect that the more traffic allowed into a network the higher the throughput. In congestion-prone systems this need not be the case, and the throughput may actually decrease as the input traffic increases (see, for example, Figure 2 of [Klei78a]). Various flow control policies have been designed to counter such behavior. However, in the queueing models we consider in this dissertation, not only is delay an increasing function of the input traffic, but also throughput is an increasing function of the traffic.

A performance measure combining throughput and delay into a single function is the notion of *power* introduced in [Gies78]. This is simply defined as

$$P_G \triangleq \frac{\gamma}{T(\gamma)} \tag{2.1}$$

where γ is the throughput and T is the mean delay as defined in chapter 1. The two contrasting objectives of maximizing throughput and minimizing delay (or maximizing $1/T$) are incorporated into this single objective function. A similar measure (P_N) was independently defined in [Yosh77], while a third related power function (P_K) appeared in [Klei79]. For the simple networks studied in this chapter, all three definitions of power yield the same optimal power point (this will be shown below in chapter 4, where we present these other definitions). Consequently, only the measure P_G (which we denote by P for simplicity) will be studied, with the understanding that the results which we obtain are equally applicable if the other measures are used. However, in subsequent chapters of this dissertation all three measures will be compared and contrasted for more general network problems (for which the optimal power point *will* depend on the particular choice of function).

Let us introduce the notion of optimizing power for a network configuration by studying the throughput-delay profile as shown in Figure 2.1.

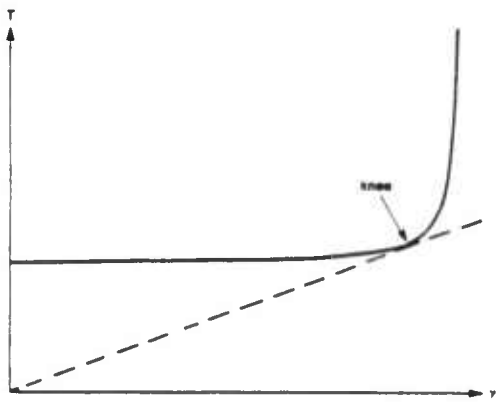

Figure 2.1 A Throughput-Delay Profile

Note that for small values of γ, a significant increase in throughput can be obtained with only a slight increase in delay, motivating us to increase throughput in this region; conversely, for large values of γ, a large decrease in delay will occur if the throughput is decreased only slightly; motivating us to decrease throughput in this region. Clearly we should avoid operating near these extreme regions; but where should we operate? If we have a system with the profile of Figure 2.1, then it is not difficult to see that an appropriate operating point for this system would be at (or near) the *"knee"* of the throughput-delay curve. Here the knee is fairly well-defined, and, as we shall see shortly, a useful choice is to define it exactly as the point on the curve such that a line through the origin to the point is tangent to the curve.

Now consider the profile shown in Figure 2.2 (which happens to be that for M/M/1).

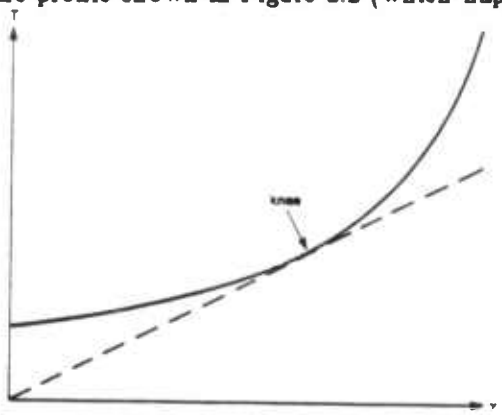

Figure 2.2 Throughput-Delay Profile for M/M/1

In this case, it is not as obvious where one should operate; to resolve this uncertainty, we again choose the knee (as defined above) as our operating point. Kleinrock [Klei78a] demonstrated the usefulness of this "knee criterion" by observing that the value of γ which maximizes power occurs exactly at the knee. We see then that an appropriate operating point for the system is to choose that value of γ which maximizes power; we now reproduce his argument.

In Figure 2.3 below, we note that a straight line through the origin to a point (γ_0, T_0) on the throughput-delay curve has slope T_0/γ_0, which is the reciprocal of the power function evaluated at that point. Thus maximizing power is equivalent to minimizing such a slope.

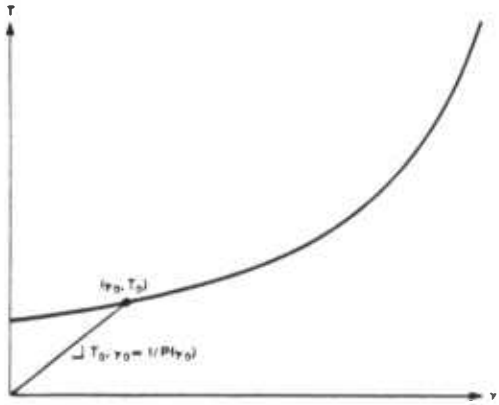

Figure 2.3 Relationship Between Slope and Power

The minimal slope occurs where a line through the origin first hits the throughput delay curve, starting with slope zero and increasing in slope until the line intersects the curve. The minimal slope clearly occurs at the knee of the throughput delay curve, and thus power is maximized at the knee. Note also that the line from the origin is tangent to the throughput delay curve at the knee, and so it must have slope $dT/d\gamma$. Equating the two expressions for the slope, we have

$$\frac{dT}{d\gamma} = \frac{T}{\gamma}$$

This shows that power is maximized for a value of γ where

$$\gamma \frac{dT}{d\gamma} = T$$

for the M/M/1 system. If we use the convention that the superscript * indicates that the variable in question has been *maximized* with respect to power, we may rewrite the above equation as

$$\gamma^* \left.\frac{dT}{d\gamma}\right|_{\gamma=\gamma^*} = T^* \tag{2.2}$$

Multiplying equation (2.2) by γ^* and using Little's result [Litt61], we obtain the equivalent equation

$$(\gamma^*)^2 \left.\frac{dT}{d\gamma}\right|_{\gamma=\gamma^*} = \bar{N}^* \tag{2.3}$$

which characterizes the maximal power point. Although this approach can be immediately generalized to *networks* which have similar throughput delay curves, Kleinrock further extended this argument by noting that the $T(\gamma)$ curve need not be a convex function of γ in order for the above to hold. In the cases where more than one tangent line can be found (more than one knee occurs), maximum power will occur for that tangent line which makes the smallest angle with the horizontal axis. Examples of nonconvex throughput delay curves may occur in the case of multiple access protocols which adapt to increasing load. This type of behavior was observed for the URN scheme of Yemini and Kleinrock [Klei78] and its extensions [Mitt81]. Although such a curve has multiple knees, equation (2.2) may still be used to determine the maximal power point.

The above geometric argument may also be applied to throughput-delay curves which are continuous, but are not necessarily differentiable everywhere (i.e., they have "cusps"). Arguing as before, we still see by the definition of P that maximizing power is equivalent to minimizing the slope of a line through the origin to the (γ, T) curve. Thus the optimal power point occurs where a line through the origin (starting with slope zero and increasing in slope) first touches the throughput-delay curve. Note, however, that the line is not necessarily tangent to the curve at the optimal power point, since the derivative $dT/d\gamma$ may not exist there, and thus equations (2.2) and (2.3) may not characterize the optimal power point.

Equation (2.2) (and thus equation (2.3)) may also be derived from optimization theory considerations [Klei78a]. The feasible region over which the power function is defined for M/M/1 is simply the closed interval of the real line with endpoints $\gamma = 0$ and $\gamma = 1/\bar{x}$ (which correspond to $\rho = 0$ and $\rho = 1$). We observe that the power function $P = \gamma/T$ is positive throughout the interior of this feasible region and zero at the two endpoints, and thus the maximum of P must occur in the interior of the feasible region. Since P is differentiable and the maximum is at an interior point, the maximum must occur when the derivative $dP/d\gamma$ is zero. But

$$\frac{dP}{d\gamma} = \frac{T - \gamma\, dT/d\gamma}{T^2}$$

and so, setting $dP/d\gamma = 0$, we see that equation (2.2) characterizes the maximal power point as before. Note that the above argument does not require any concavity assumptions on P; it will hold for any differentiable power function which is zero at the endpoints of the feasible region and positive in the interior. However, if the function does not have "nice" properties such as concavity, there may be several points where the derivative is zero. In this case, equation (2.2) will hold at each such point, and these points must be compared to determine the global maximum.

Using equation (2.2), Kleinrock [Klei79] showed the remarkable result that for any M/G/1 system

$$\bar{N}^* = 1 \tag{2.4}$$

Thus equation (2.4) indicates that if the throughput γ (or equivalently, the input rate λ) of an M/G/1 system is chosen so as to maximize power $P(\gamma)$, then the average number in system $\bar{N}^*$ at this point γ^* is equal to 1. This agrees with the reasoning that the proper operating point of the *deterministic* D/D/1 system is exactly when $\bar{N}^* = 1$. It also can be shown that the server utilization at maximum power for M/G/1 is

$$\rho^* = \frac{1}{1 + \sqrt{(1+\nu^2)/2}} \tag{2.5}$$

where ν is the coefficient of variation of the service time distribution (recall that ν is simply the ratio of the standard deviation to the mean service time). Other networks, such as tandems, have also been analyzed from the point of view of maximizing power, and various results have appeared in the literature [Klei78a, Klei79, Bhar80]. We will now extend the analysis of power to more general "simple" computer network configurations. In each case, the optimization problem we study involves a single decision variable. The feasible region is a closed (and bounded) interval of the real line, and the power function is zero at the two endpoints (which correspond to zero throughput and infinite delay). Thus the point which maximizes power occurs in the interior of the feasible region, and so equations (2.2) and (2.3) characterize this optimal power point.

2.2 The M/M/1 Series Network

A message path in a network can be modeled as a series of independent M/M/1 queues, if there is sufficient mixing of messages to justify the use of Kleinrock's independence assumption [Klei64]. This M/M/1 series network has been analyzed with respect to power by Kleinrock [Klei78a] and Bharath-Kumar [Bhar80]. In this section, we reproduce several of their results for continuity of exposition, and we also obtain other results for the M/M/1 tandem. We consider a tandem of M channels where the ith channel has capacity C_i bits per second (see Figure 2.4).

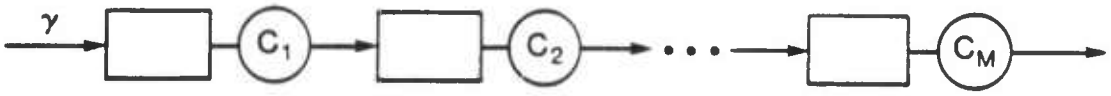

Figure 2.4 The M/M/1 Series Network

We assume that the arrival process of messages to the system is Poisson with rate λ, and that the message length distribution is exponential with mean $\bar{b} = 1/\mu$ bits. Note that the throughput of the system γ is equal to the input rate λ. We also invoke the independence assumption in order to view each channel as an M/M/1 queueing system. Kleinrock [Klei78a] showed that for such a series with M channels and *equal* channel capacities ($C_i \triangleq C$ for $1 \leq i \leq M$), then

$$\bar{N}^* = M \tag{2.6}$$

In fact, he showed that the optimal value of throughput is $\gamma^* = 1/(2\bar{x}) = \mu C/2$, and it therefore occurs when $\rho_i^* = 1/2$. Thus

$$\bar{N}_i^* = 1 \tag{2.7}$$

for $1 \leq i \leq M$. This model of a message path in a computer network was further examined by Bharath-Kumar [Bhar80]. He considered such a path having *arbitrary* channel capacities, and found that

$$\bar{N}^* \leq M \tag{2.8}$$

with Kleinrock's "keep the pipe full" result of equation (2.6) holding when all capacities are equal.

We first give a proof of Bharath-Kumar's bound of equation (2.8), since an expression derived in the course of the argument will be of use later. Maximizing power for the M/M/1 series network yields a simple (single variable) optimization problem. The endpoints of the feasible region are $\gamma = 0$ and $\gamma = \mu C_{min}$, where C_{min} is the smallest of the M channel capacities in the tandem. Note that P is zero at both endpoints, and thus we may use equation (2.3) to find the optimal power point for this system. The average time that a message spends in the system T is simply

$$T = \sum_{i=1}^{M} T_i$$

where T_i is the mean time a message spends at the ith channel (both in queue and in transmission). Since each channel acts like an M/M/1 queueing system, we have

$$T_i = \frac{\bar{x}_i}{1 - \rho_i}$$

where $\rho_i = \gamma \bar{x}_i$ is the utilization of the ith channel (and $\bar{x}_i = 1/\mu C_i$). Now note that

$$\frac{dT_i}{d\gamma} = \left(\frac{\bar{x}_i}{1 - \rho_i}\right)^2 = (T_i)^2$$

so that

$$\frac{dT}{d\gamma} = \sum_{i=1}^{M} \frac{dT_i}{d\gamma} = \sum_{i=1}^{M} (T_i)^2$$

Substituting this relationship into equation (2.3) yields

$$\bar{N}^* = (\gamma^*)^2 \sum_{i=1}^{M} (T_i^*)^2$$

Using Little's result, we obtain the following theorem proved by Bharath-Kumar for the M/M/1 series network.

Theorem 2.1 (Bharath-Kumar)

For the M/M/1 series network, the average number in system at maximum power satisfies

$$\bar{N}^* = \sum_{i=1}^{M} \bar{N}_i^* = \sum_{i=1}^{M} (\bar{N}_i^*)^2 \tag{2.9}$$

Note that if all channel capacities are equal, then the $\bar{N}_i^*$ are identical for all i, and Kleinrock's equations (2.6) and (2.7) follow. From equation (2.9), Bharath-Kumar derived the upper bound given in equation (2.8). We now give a simpler proof of his result. Equation (2.9) yields

$$\overline{N}^* = \sum_{i=1}^{M} \overline{N}_i^* = \sum_{i=1}^{M} \left[\overline{N}_i^* + \left[\overline{N}_i^* - (\overline{N}_i^*)^2 \right] \right]$$

The above expression is equivalent to

$$\overline{N}^* = \sum_{i=1}^{M} \left[1 - 1 + 2\overline{N}_i^* - (\overline{N}_i^*)^2 \right]$$

or

$$\overline{N}^* = M - \sum_{i=1}^{M} (1 - \overline{N}_i^*)^2$$

This immediately gives equation (2.8).

We will now find a *lower* bound on the average number in system at maximum power for the M/M/1 series network with arbitrary channel capacities. We will show that

$$\sum_{i=1}^{M} \frac{C_{\min}}{C_i} \leq \overline{N}^* \tag{2.10}$$

where $C_{\min}$ is the minimum of the M channel capacities of the tandem. To state this lower bound in another way, let us express $C_i = \alpha_i C_{\min}$ for each i (where, of course, $\alpha_i \geq 1$). Thus we wish to show

$$\sum_{i=1}^{M} \frac{1}{\alpha_i} \leq \overline{N}^* \tag{2.11}$$

To obtain the lower bound, we will use Bharath-Kumar's equation (2.9). We may write

$$T_i = \overline{x}_i + W_i = \overline{x}_i + \frac{\rho_i \overline{x}_i}{1 - \rho_i}$$

since node i is M/M/1. Multiplying by γ and using Little's result yields

$$\overline{N}_i = \rho_i + \frac{(\rho_i)^2}{1 - \rho_i} = \rho_i + \left(\frac{\rho_i}{1 - \rho_i} \right)^2 (1 - \rho_i)$$

or

$$\overline{N}_i = \rho_i + (\overline{N}_i)^2 (1 - \rho_i)$$

Thus

$$\overline{N} = \sum_{i=1}^{M} \overline{N}_i = \sum_{i=1}^{M} \rho_i + \sum_{i=1}^{M} (\overline{N}_i)^2 (1 - \rho_i)$$

or

$$\overline{N} = \sum_{i=1}^{M} \rho_i + \sum_{i=1}^{M} (\overline{N}_i)^2 - \sum_{i=1}^{M} \rho_i (\overline{N}_i)^2$$

This equation holds for all values of throughput γ. If we evaluate it at the maximum power point and use equation (2.9) to simplify the resulting expression, we have

$$\sum_{i=1}^{M} \rho_i^* = \sum_{i=1}^{M} \rho_i^* (\overline{N}_i^*)^2$$

Dividing both sides of this equation by ρ_{max}^* yields

$$\sum_{i=1}^{M} \frac{\rho_i^*}{\rho_{max}^*} = \sum_{i=1}^{M} \frac{\rho_i^*}{\rho_{max}^*} (\overline{N}_i^*)^2$$

Now since $\rho_i^*/\rho_{max}^* \leq 1$ for each i, we have

$$\sum_{i=1}^{M} \frac{\rho_i^*}{\rho_{max}^*} = \sum_{i=1}^{M} \frac{\rho_i^*}{\rho_{max}^*} (\overline{N}_i^*)^2 \leq \sum_{i=1}^{M} (\overline{N}_i^*)^2$$

Thus using equation (2.9) again, we obtain

$$\sum_{i=1}^{M} \frac{\rho_i^*}{\rho_{max}^*} \leq \overline{N}^* \tag{2.12}$$

We now write this lower bound entirely in terms of the (given) channel capacities. Noting that $\rho_i^* = (\gamma^* \overline{b})/C_i$ and thus $\rho_{max}^* = (\gamma^* \overline{b})/C_{min}$ we obtain equation (2.10), namely,

$$\sum_{i=1}^{M} \frac{C_{min}}{C_i} \leq \overline{N}^*$$

Rewriting this equation yields equation (2.11), that is,

$$\sum_{i=1}^{M} \frac{1}{\alpha_i} \leq \overline{N}^*$$

Therefore, the lower bound of equation (2.10) which we have just shown and the upper bound of equation (2.8) due to Bharath-Kumar yield the following

Theorem 2.2

For the M/M/1 series network, the average number in system at maximum power satisfies

$$\sum_{i=1}^{M} \frac{C_{min}}{C_i} \leq \overline{N}^* \leq M \tag{2.13}$$

Note that the upper and lower bounds are equal if we have equal capacities at all M nodes; in this case we obtain Kleinrock's result that $\overline{N}^* = M$. However, we also note that the upper and lower bounds of Theorem 2.2 may be quite different in some cases. For example, we consider a tandem with one slow channel (of capacity C) and $M-1$ fast channels (of capacity βC, where $\beta >> M$). Here the lower bound is $\sum_{i=1}^{M} \frac{C_{min}}{C_i} = 1 + \frac{M-1}{\beta} \cong 1$ which is far less than the upper bound of M (especially if M is large). Since the throughput is constrained by the

one bottleneck node, the average number in system at optimal power is approximately 1. Thus the lower bound is close to the actual value of $\overline{N}^*$, while the upper bound is quite bad.

We now turn our attention to the optimal throughput γ^*. Since

$$\overline{N}_i^* = \frac{\rho_i^*}{1-\rho_i^*} = \frac{\gamma^*}{\mu C_i - \gamma^*}$$

for $1 \leq i \leq M$, equation (2.9) can be written solely in terms of γ^* as

$$\sum_{i=1}^{M} \frac{\gamma^*}{\mu C_i - \gamma^*} = \sum_{i=1}^{M} \left[\frac{\gamma^*}{\mu C_i - \gamma^*}\right]^2$$

This yields a polynomial equation in γ^*, and thus, as observed by Bharath-Kumar [Bhar80], γ^* may be found by any polynomial root finding procedure. Bharath-Kumar also found upper and lower bounds on the value of γ^* in [Bhar80] for the M/M/1 tandem. The upper bound he gives is simply $\gamma^* < \mu C_{\min}$, which is the upper endpoint of the feasible region. This must hold in order for the tandem to remain stable. The lower bound he gives is incorrect. We now show upper and lower bounds on γ^* by a method completely different from that of Bharath-Kumar. However, a proper utilization of his method will also give these bounds. We first need to show the following.

Lemma: Let f be a real-valued differentiable strictly convex function defined on an open convex subset S of the real numbers. Let $x, y \in S$ with $0 \leq x < y$. Then

$$f(y) - f(x) < f'(y)\, y - f'(x)\, x \tag{2.14}$$

This is simple to prove as follows. Since $x < y$ (and thus $x \neq y$), it is well-known (see Theorem 6.2.3 of [Mang69]) that

$$[f'(y) - f'(x)]\,(y - x) > 0$$

and so $f'(x) < f'(y)$. Also we have (see Theorem 6.2.2 of [Mang69]) that

$$f(x) - f(y) > f'(y)\,(x - y)$$

or, equivalently,

$$f(y) - f(x) < f'(y)\,(y - x)$$

Thus, since $f'(y)\, x \geq f'(x)\, x$,

$$f(y) - f(x) < f'(y)\, y - f'(y)\, x \leq f'(y)\, y - f'(x)\, x$$

which is equation (2.14).

Using the lemma we will now find upper and lower bounds on the value of γ^* by proving the following

Theorem 2.3

For the M/M/1 series network the value of throughput which maximizes power satisfies

$$\frac{\mu C_{\min}}{2} \leq \gamma^* \leq \frac{\mu C_{\max}}{2} \tag{2.15}$$

(Of course we also have $\gamma^* < \mu C_{\min}$ in order for the system to be stable.) Note that equal capacities at all M nodes gives Kleinrock's result that $\gamma^* = \mu C/2$.

We now prove Theorem 2.3. Let λ_i^* maximize power for the ith channel individually, and recall that γ^* maximizes power for the M channel tandem. We now show

$$\min_{1 \leq i \leq M} \lambda_i^* \leq \gamma^* \leq \max_{1 \leq i \leq M} \lambda_i^* \tag{2.16}$$

Since each individual node is an M/M/1 system, $\lambda_i^* = \mu C_i/2$. Thus equation (2.16) implies equation (2.15), and so we need only prove equation (2.16). We first show the left-hand inequality of equation (2.16). To this end, suppose (for proof by contradiction) that

$$\gamma^* < \min_{1 \leq i \leq M} \lambda_i^*$$

Then

$$\gamma^* < \lambda_i^*$$

for $1 \leq i \leq M$. We now note that the average delay T_i at channel i is a differentiable strictly convex function of γ in the open set $S_i \triangleq (0, \mu C_i)$. Also note that $\gamma^* < \mu C_{\min} \leq \mu C_i$, so that $\lambda_i^*, \gamma^* \in S_i$. Thus the above lemma (equation (2.14)) may be used. By this previous lemma applied to T_i we have

$$T_i(\lambda_i^*) - T_i(\gamma^*) < \lambda_i^* \frac{dT_i}{d\gamma}\Big|_{\gamma=\lambda_i^*} - \gamma^* \frac{dT_i}{d\gamma}\Big|_{\gamma=\gamma^*}$$

for $1 \leq i \leq M$. From the optimality of λ_i^*, equation (2.2) yields

$$\lambda_i^* \frac{dT_i}{d\gamma}\Big|_{\gamma=\lambda_i^*} = T_i(\lambda_i^*)$$

for $1 \leq i \leq M$. Therefore

$$\gamma^* \frac{dT_i}{d\gamma}\Big|_{\gamma=\gamma^*} < T_i(\gamma^*)$$

for $1 \leq i \leq M$. Adding these M inequalities and recognizing $T = \sum_{i=1}^{M} T_i$ so that $\frac{dT}{d\gamma} = \sum_{i=1}^{M} \frac{dT_i}{d\gamma}$, we have

$$\gamma^* \left.\frac{dT}{d\gamma}\right|_{\gamma=\gamma^*} < T(\gamma^*)$$

But since γ^* maximizes power for the M channel tandem, equation (2.2) gives

$$\gamma^* \left.\frac{dT}{d\gamma}\right|_{\gamma=\gamma^*} = T(\gamma^*)$$

This contradiction shows that

$$\min_{1 \le i \le M} \lambda_i^* \le \gamma^*$$

Similarly we may show the right-hand inequality of equation (2.16), namely

$$\gamma^* \le \max_{1 \le i \le M} \lambda_i^*$$

Thus equation (2.16) (and therefore Theorem 2.3) is proved.

From the above results we already see invariant properties of $\overline{N}^*$, the average number in system at maximum power. For example, equation (2.4) shows an invariance with respect to service time distribution, while equation (2.13) shows an invariance with respect to scaling of channel capacities. We know of no other system variable which shows such invariance. In the following sections we introduce other simple networks and again emphasize such invariance properties in our analyses.

2.3 The M/D/1 Series Network

Consider a computer network with constant length messages of $\overline{b}$ bits. Furthermore, as a message traverses a path through the network, its message length does not change. Its service time (or transmission time) over a channel of capacity C bits per second is simply $\overline{x} = \overline{b}/C$. We choose to model a message path in a computer network as a series of M queues where the length of a message remains *constant* as it traverses the path. Note that the arrival rate λ and the throughput γ are identical for this series network. Also note that the various channels may, in general, have different capacities C_i (see Figure 2.5).

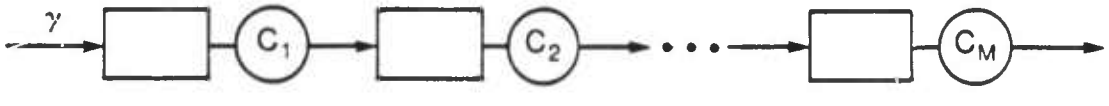

Figure 2.5 The M/D/1 Series Network

If we assume *Poisson* arrivals, the first node in the system is simply an M/D/1 queue. However, the subsequent nodes in the path are more complicated queueing systems (there is a dependence between the output process of one node and the input process of the next node). This model was extensively analyzed by Rubin [Rubi74]. He showed that the distribution of waiting time for the total system is identical to that for an M/D/1 queueing system with arrival rate λ and service time equal to the *maximum* of the individual service times at the nodes. With this result, the Pollaczek-Khinchin equation [Klei75] may be used to give the mean end-to-end waiting time for a message along the path as

$$W = \frac{\gamma \overline{x^2}_{\max}}{2(1-\rho_{\max})} = \frac{\gamma(\overline{x}_{\max})^2}{2(1-\rho_{\max})} \tag{2.17}$$

and so

$$\gamma W = \frac{(\rho_{\max})^2}{2(1-\rho_{\max})} \tag{2.18}$$

This expression gives the average delay in the system as

$$T = W + \sum_{i=1}^{M} \overline{x}_i = \frac{\gamma(\overline{x}_{\max})^2}{2(1-\gamma\overline{x}_{\max})} + \sum_{i=1}^{M} \overline{x}_i \tag{2.19}$$

where M is the number of channels in the tandem. Differentiating this equation we have

$$\frac{dT}{d\gamma} = \frac{dW}{d\gamma} = \frac{(\overline{x}_{\max})^2}{2(1-\gamma\overline{x}_{\max})^2} \tag{2.20}$$

Using equations (2.3) and (2.20), at maximum power we have

$$\overline{N}^* = (\gamma^*)^2 \left.\frac{dT}{d\gamma}\right|_{\gamma=\gamma^*} = \frac{(\rho^*_{\max})^2}{2(1-\rho^*_{\max})^2} \tag{2.21}$$

From equation (2.18), we obtain

$$\overline{N}^* = \frac{\gamma^* W^*}{1-\rho^*_{\max}} = \frac{\overline{N}^* - \sum_{i=1}^{M} \rho_i^*}{1-\rho^*_{\max}}$$

or

$$\overline{N}^* - \rho^*_{\max}\overline{N}^* = \overline{N}^* - \sum_{i=1}^{M} \rho_i^*$$

This last equation yields

$$\overline{N}^* = \sum_{i=1}^{M} \frac{\rho_i^*}{\rho^*_{\max}} \tag{2.22}$$

Noting that $\rho_i^* = (\gamma^*\overline{b})/C_i$ and thus $\rho^*_{\max} = (\gamma^*\overline{b})/C_{\min}$ we obtain

Theorem 2.4

For the M/D/1 series network, power is maximized when

$$\overline{N}^* = \sum_{i=1}^{M} \frac{C_{\min}}{C_i} \tag{2.23}$$

Note that this expression is solely in terms of the given quantities C_i. It does not involve γ^* and is invariant under scaling of the channel capacities.

Let us express this important equation in another way which will lead to an interpretation which is intuitively pleasing. For each node i $(1 \leq i \leq M)$ set $C_i = \alpha_i C_{\min}$ where $\alpha_i \geq 1$. Then

$$\sum_{i=1}^{M} \frac{C_{\min}}{C_i} = \sum_{i=1}^{M} \frac{1}{\alpha_i}$$

and so we may write Theorem 2.4 in the form

$$\overline{N}^* = \sum_{i=1}^{M} \frac{1}{\alpha_i} \tag{2.24}$$

Thus the slowest channel (capacity $C_{\min}$) contributes 1 to the average number in system while a channel which is $\alpha \geq 1$ times faster than the slowest channel contributes $1/\alpha \leq 1$ to this average number. This is similar to the "keep the pipe full" result of Kleinrock mentioned above. However, although equation (2.24) holds, $\overline{N}_i^* \neq 1/\alpha_i$ in general (see, for example, the case of equal channel capacities discussed below).

Theorem 2.4 shows that the average number in system at maximum power does not depend on the *order* of the individual capacities C_i, but only on their values. If the first channel is the slowest, no queueing takes place except at node 1; if the last channel is the slowest, queueing may take place at various nodes. Thus the *individual* values of $\overline{N}_i^*$ for $1 \leq i \leq M$ *do* depend on the order of the capacities. But in all cases we have the invariance for the *total* number, i.e.,

$$\overline{N}^* = \sum_{i=1}^{M} \frac{1}{\alpha_i}$$

Clearly we have the same upper bound as for the M/M/1 series network, namely,

$$\overline{N}^* \leq M \tag{2.25}$$

Recall that, for the M/M/1 series network, Theorem 2.3 gives the bound

$$\sum_{i=1}^{M} \frac{1}{\alpha_i} \leq \overline{N}^*_{M/M/1}$$

Thus we have the interesting result that

$$\bar{N}^*_{M/D/1} \leq \bar{N}^*_{M/M/1} \tag{2.26}$$

Of course we have equality (both means have value M) if all capacities are equal.

As mentioned earlier, the M/D/1 series network may be used to model a virtual circuit in a packet network. A window flow control scheme is often used with virtual circuits. Our result that $\bar{N}^* = \sum_{i=1}^{M} \frac{C_{min}}{C_i}$ (equation (2.23)) may therefore be used as the window size setting for this path. This number is easily calculable, and uses only local path information in the sense of Jaffe [Jaff81].

The above M/D/1 expression for $\bar{N}^*$ may also be used as an *approximation* for the average number in system at maximum power of the M/M/1 series model with arbitrary capacities. As explained in [Bhar80] the calculation of $\bar{N}^*$ for the M/M/1 tandem involves the complexity of finding roots of polynomials. We have numerically calculated values for average number in system for the M/M/1 series model with up to 100 channels and various combinations of capacities. We have compared the results to values obtained by operating the M/M/1 tandem at a window size given from the M/D/1 equation (2.23). The error in value of power was found to be small, although other system parameters had larger errors in some cases. The best cases were those with fewer channels (as one might expect) and those with one slow server and all the rest fast (the bottleneck case). This is pleasing because the bound $\bar{N}^* \leq M$ for M/M/1 given in [Bhar80] (and rederived as part of Theorem 2.2 above) can be quite bad. Recall the example M/M/1 series network discussed after Theorem 2.2 which has one slow channel of capacity C and $M-1$ channels of capacity βC where $\beta >> M$. This tandem has a value for $\bar{N}^*$, the average number in system at optimal power, which is close to 1 (far from the upper bound of M). But the M/D/1 value is $\sum_{i=1}^{M} \frac{1}{\alpha_i} = 1 + \frac{M-1}{\beta} \cong 1$ as expected. Thus equation (2.23) may often be used as a good approximation to the M/M/1 tandem.

Other optimized variables of interest for the M/D/1 series net include γ^*, T^*, P^* and ρ_i^* for $1 \leq i \leq M$. By equation (2.21) we have

$$\bar{N}^* = \frac{(\rho^*_{max})^2}{2(1-\rho^*_{max})^2}$$

and so we obtain the quadratic equation

$$(\rho^*_{max})^2 = 2\bar{N}^*(1-\rho^*_{max})^2$$

Taking square roots gives

$$\rho^*_{max} = \sqrt{2\overline{N}^*}(1 - \rho^*_{max})$$

and so

$$\rho^*_{max} = \frac{\sqrt{2\overline{N}^*}}{1+\sqrt{2\overline{N}^*}} \tag{2.27}$$

Other variables, such as ρ_i^* for $1 \leq i \leq M$, and γ^*, T^*, $P^* = P(\gamma^*)$, may also be calculated easily from our earlier equations. Recall that $\rho_i^* = (\gamma^* \overline{b})/C_i$ and so $\rho^*_{max} = (\gamma^* \overline{b})/C_{min}$. Therefore

$$\rho_i^* = \frac{\rho_i^*}{\rho^*_{max}} \cdot \rho^*_{max} = \frac{C_{min}}{C_i}\left[\frac{\sqrt{2\overline{N}^*}}{1+\sqrt{2\overline{N}^*}}\right]$$

for $1 \leq i \leq M$. The optimal throughput is

$$\gamma^* = \frac{\rho^*_{max}}{\overline{b}/C_{min}} = \frac{C_{min}}{\overline{b}}\left[\frac{\sqrt{2\overline{N}^*}}{1+\sqrt{2\overline{N}^*}}\right]$$

We also have

$$T^* = \frac{\overline{N}^*}{\gamma^*} = \frac{\overline{b}/C_{min}}{\rho^*_{max}}\overline{N}^* = \frac{\overline{b}}{C_{min}}\left[\frac{\sqrt{2\overline{N}^*}+2\overline{N}^*}{2}\right]$$

and

$$P^* = P(\gamma^*) = \frac{\gamma^*}{T^*} = \frac{(\gamma^*)^2}{\overline{N}^*} = \left[\frac{C_{min}}{\overline{b}}\right]^2 \cdot \frac{2}{\left[1+\sqrt{2\overline{N}^*}\right]^2}$$

For the case of equal channel capacities, we obtain a simplification of the above results. Assuming $C_i \triangleq C$ for $1 \leq i \leq M$ and thus $\rho_i \triangleq \rho$, Theorem 2.4 gives

$$\overline{N}^* = M \tag{2.28}$$

This is the same result as Kleinrock [Klei78a] obtained for the M/M/1 tandem with equal capacities (see equation (2.6)). However, the average numbers at the individual nodes at optimal power $\overline{N}_i^*$ are different for the two series networks. In the M/D/1 tandem with equal channel capacities, all queueing takes place at the first channel; there is no waiting at any of the subsequent channels. Therefore, $\overline{N}_1^* \neq \overline{N}_i^*$ for $2 \leq i \leq M$. In the M/M/1 tandem, Kleinrock showed that $\overline{N}_i^* = 1$ for $1 \leq i \leq M$ (see equation (2.7)). Therefore, the individual average numbers $\overline{N}_i^*$ are identical for $1 \leq i \leq M$.

For the M/D/1 series network with equal channel capacities, equation (2.27) yields

$$\rho^* = \frac{\sqrt{2M}}{1+\sqrt{2M}} \tag{2.29}$$

which compares with the value of $\rho^* = 1/2$ mentioned above for the M/M/1 series net with equal channel capacities. Thus values of other variables (γ^*, ρ^*, T^*) differ between the M/M/1 and M/D/1 series models as do the individual node average numbers $\bar{N}_i^*$ for $1 \leq i \leq M$. It is all the more amazing that the expression $\bar{N}^* = M$ is invariant for both service time *distributions*. Table 2.1 shows the differences in the values of various optimized variables for the two systems.

	M/M/1			M/D/1		
	Average Number In			Average Number In		
Node	Queue	Server	System	Queue	Server	System
1	$1-\rho^*$	ρ^*	1	$M(1-\rho^*)$	ρ^*	$M(1-\rho^*)+\rho^*$
2	$1-\rho^*$	ρ^*	1	0	ρ^*	ρ^*
3	$1-\rho^*$	ρ^*	1	0	ρ^*	ρ^*
⋮	⋮	⋮	⋮	⋮	⋮	⋮
M	$1-\rho^*$	ρ^*	1	0	ρ^*	ρ^*
Total	$M(1-\rho^*)$	$M\rho^*$	M	$M(1-\rho^*)$	$M\rho^*$	M

Table 2.1

This table graphically illustrates that, although $\bar{N}^* = M$ for both series networks, the behavior of the two systems at optimal power is quite different. Noting that $\rho^* = 1/2$ for the M/M/1 series network, and that $\rho^* = \sqrt{2M}/[1+\sqrt{2M}]$ for the series network with constant message length, we obtain the optimized values shown in Table 2.2.

	M/M/1			M/D/1		
	Average Number In			Average Number In		
Node	Queue	Server	System	Queue	Server	System
1	$\frac{1}{2}$	$\frac{1}{2}$	1	$\frac{M}{1+\sqrt{2M}}$	$\frac{\sqrt{2M}}{1+\sqrt{2M}}$	$\frac{M+\sqrt{2M}}{1+\sqrt{2M}}$
2	$\frac{1}{2}$	$\frac{1}{2}$	1	0	$\frac{\sqrt{2M}}{1+\sqrt{2M}}$	$\frac{\sqrt{2M}}{1+\sqrt{2M}}$
3	$\frac{1}{2}$	$\frac{1}{2}$	1	0	$\frac{\sqrt{2M}}{1+\sqrt{2M}}$	$\frac{\sqrt{2M}}{1+\sqrt{2M}}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮
M	$\frac{1}{2}$	$\frac{1}{2}$	1	0	$\frac{\sqrt{2M}}{1+\sqrt{2M}}$	$\frac{\sqrt{2M}}{1+\sqrt{2M}}$
Total	$\frac{M}{2}$	$\frac{M}{2}$	M	$\frac{M}{1+\sqrt{2M}}$	$\frac{M\sqrt{2M}}{1+\sqrt{2M}}$	M

Table 2.2

We see from Table 2.2 that, in the M/M/1 tandem, each of the M channels contributes exactly 1 to the average number (for any value of M). In the M/D/1 series net with equal capacities there is no queueing at any node i for $2 \leq i \leq M$ (all queueing occurs at the first channel). In fact, for the M/D/1 tandem, as $M \to \infty$ then the common utilization $\rho^* \to 1$ (almost 1 in each server) and also the number in queue at the first node $M(1-\rho^*) \to \infty$. As the table indicates, other system variables (ρ^*, $\overline{N}_i^*$, etc.) differ greatly for the two models, but in both cases amazingly $\overline{N}^* = M$.

2.4 The M/G/1 Parallel Network (Known Routing)

This network model is for a parallel configuration of M message channels, where channel i has capacity C_i. The input to the system is again Poisson with parameter λ. The system throughput γ is equal to the input rate λ. The message length distribution is *arbitrary*. In particular, the length of messages which use channel i is a random variable $\tilde{b}_i$ (with arbitrary distribution), and so the service time (transmission time) for a message on the ith channel is the random variable $\tilde{x}_i = \tilde{b}_i/C_i$. Probabilities $0 < p_i < 1$ for $1 \leq i \leq M$ are given satisfying

$\sum_{i=1}^{M} p_i = 1$ which determine the channel a newly arrived message chooses. Thus each channel i acts as an M/G/1 queueing system with Poisson input rate $\lambda_i = p_i \lambda = p_i \gamma$ for $1 \leq i \leq M$ (see Figure 2.6).

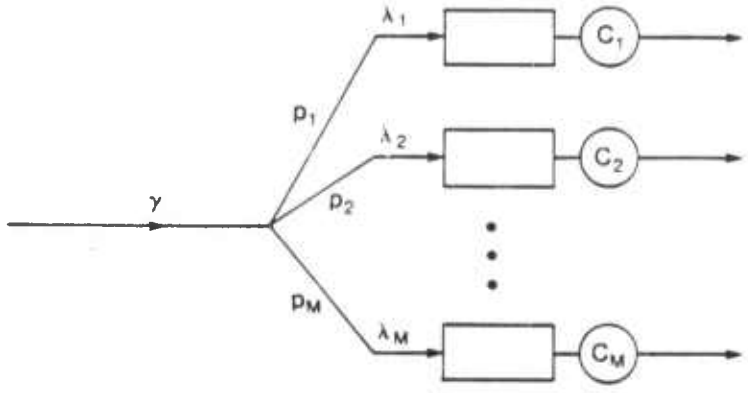

Figure 2.6 The M/G/1 Parallel Network

We wish to find that value of γ which will maximize the power of this parallel system. Since each node is an M/G/1 system, then by th Pollaczek–Khinchin equation [Klei75]

$$W_i = \frac{\lambda_i \overline{x_i^2}}{2(1-\rho_i)} = \frac{p_i \gamma \overline{x_i^2}}{2(1-p_i \gamma \bar{x}_i)} \tag{2.30}$$

where $\rho_i = \lambda_i \bar{x}_i = p_i \gamma \bar{x}_i$ is the utilization of the ith channel, $\bar{x}_i = \bar{b}_i / C_i$ is the mean service time for channel i, and $\overline{x_i^2} = \overline{b_i^2}/C_i^2$ is the second moment of service time for channel i. Therefore

$$\lambda_i W_i = \frac{(\lambda_i)^2 \overline{x_i^2}}{2(1-\rho_i)} \tag{2.31}$$

We may also write W_i in terms of ν_i, the coefficient of variation of service time for channel i (i.e., the ratio of the standard deviation to the mean service time). Note that ν_i is simply equal to the coefficient of variation of message length for the ith channel. Equation (2.30) becomes

$$W_i = \frac{\rho_i \bar{x}_i}{1-\rho_i}\left(\frac{1+\nu_i^2}{2}\right)$$

while equation (2.31) is

$$\lambda_i W_i = \frac{(\rho_i)^2}{1-\rho_i}\left(\frac{1+\nu_i^2}{2}\right) \tag{2.32}$$

We note that the average delay T_i at channel i is

$$T_i = W_i + \bar{x}_i = \frac{p_i \gamma \overline{x_i^2}}{2(1-p_i \gamma \bar{x}_i)} + \bar{x}_i$$

and the average number $\bar{N}_i$ at channel i is

$$\bar{N}_i = \lambda_i T_i = \rho_i + \lambda_i W_i \tag{2.33}$$

Writing $\bar{N}_i$ in terms of ν_i, we have

$$\bar{N}_i = \rho_i + \frac{(\rho_i)^2}{1-\rho_i}\left(\frac{1+\nu_i^2}{2}\right) \tag{2.34}$$

We may express the average delay T for the system as

$$T = \sum_{i=1}^{M} \frac{\lambda_i}{\gamma} T_i = \sum_{i=1}^{M} p_i T_i = \sum_{i=1}^{M} p_i W_i + \sum_{i=1}^{M} p_i \bar{x}_i \tag{2.35}$$

and the average number $\bar{N}$ in the system as

$$\bar{N} = \gamma T = \sum_{i=1}^{M} \lambda_i T_i = \sum_{i=1}^{M} \bar{N}_i$$

We now examine the characteristics of the optimal power point for the M/G/1 parallel network. Differentiating equation (2.35) with respect to γ yields

$$\frac{dT}{d\gamma} = \sum_{i=1}^{M} p_i \frac{dW_i}{d\gamma}$$

and thus, differentiating equation (2.30), we have

$$\frac{dT}{d\gamma} = \sum_{i=1}^{M} p_i \frac{p_i \overline{x_i^2}}{2(1-\rho_i)^2} \tag{2.36}$$

Using equations (2.3) and (2.36), at maximum power we have

$$\bar{N}^* = (\gamma^*)^2 \left.\frac{dT}{d\gamma}\right|_{\gamma=\gamma^*} = (\gamma^*)^2 \sum_{i=1}^{M} \frac{(p_i)^2 \overline{x_i^2}}{2(1-\rho_i^*)^2}$$

or

$$\bar{N}^* = \sum_{i=1}^{M} \frac{(\lambda_i^*)^2 \overline{x_i^2}}{2(1-\rho_i^*)^2} \tag{2.37}$$

Therefore, equation (2.31) yields

$$\bar{N}^* = \sum_{i=1}^{M} \frac{\lambda_i^* W_i^*}{1-\rho_i^*} \tag{2.38}$$

and thus, by equation (2.33), we have

$$\bar{N}^* = \sum_{i=1}^{M} \bar{N}_i^* = \sum_{i=1}^{M} \frac{\bar{N}_i^* - \rho_i^*}{1-\rho_i^*} \tag{2.39}$$

Rewriting the middle term of the above equation, we obtain

$$\sum_{i=1}^{M} \frac{\bar{N}_i^* - \rho_i^* \bar{N}_i^*}{1-\rho_i^*} = \sum_{i=1}^{M} \frac{\bar{N}_i^* - \rho_i^*}{1-\rho_i^*}$$

or

$$\sum_{i=1}^{M} \frac{\rho_i^*}{1-\rho_i^*}\left(1 - \bar{N}_i^*\right) = 0 \tag{2.40}$$

From equation (2.33), we also may rewrite equation (2.40) as

$$\sum_{i=1}^{M} \frac{\rho_i^*}{1-\rho_i^*}\left(1 - \rho_i^* - \lambda_i^* W_i^*\right) = 0$$

or

$$\sum_{i=1}^{M} \rho_i^* \left(1 - \frac{\lambda_i^* W_i^*}{1-\rho_i^*}\right) = 0 \tag{2.41}$$

We now use equation (2.41) to obtain several interesting results for the M/G/1 parallel network. We first find the optimal throughput γ^*. Substituting equation (2.32) in equation (2.41) yields

$$\sum_{i=1}^{M} \rho_i^* \left[1 - \left(\frac{\rho_i^*}{1-\rho_i^*}\right)^2 \left(\frac{1+\nu_i^2}{2}\right)\right] = 0 \tag{2.42}$$

Using $\rho_i^* = \lambda_i^* \bar{x}_i = \gamma^* p_i \bar{x}_i$ for $1 \le i \le M$, we see that equation (2.42) is equivalent to a polynomial equation involving the known quantities p_i, $\bar{x}_i$ and ν_i and the single unknown γ^*. Therefore, the optimal throughput γ^* may be found by any polynomial root finding procedure, and then other system variables at maximal power (such as ρ_i^* and $\bar{N}_i^*$) may be determined for the M/G/1 parallel network.

Next we use equation (2.41) to find a lower bound on $\bar{N}^*$ as follows. Rewriting equation (2.41) we have

$$\sum_{i=1}^{M} \rho_i^* = \sum_{i=1}^{M} \rho_i^* \frac{\lambda_i^* W_i^*}{1-\rho_i^*}$$

Dividing both sides of this equation by $\rho_{\max}^*$ yields

$$\sum_{i=1}^{M} \frac{\rho_i^*}{\rho_{\max}^*} = \sum_{i=1}^{M} \frac{\rho_i^*}{\rho_{\max}^*} \cdot \frac{\lambda_i^* W_i^*}{1-\rho_i^*}$$

Now since $\rho_i^*/\rho_{\max}^* \le 1$ for each i, we have

$$\sum_{i=1}^{M} \frac{\rho_i^*}{\rho_{\max}^*} = \sum_{i=1}^{M} \frac{\rho_i^*}{\rho_{\max}^*} \cdot \frac{\lambda_i^* W_i^*}{1-\rho_i^*} \le \sum_{i=1}^{M} \frac{\lambda_i^* W_i^*}{1-\rho_i^*}$$

From equation (2.38), we recognize the right-hand side of the above inequality as $\bar{N}^*$. Thus we

obtain the lower bound

$$\sum_{i=1}^{M} \frac{\rho_i^*}{\rho_{\max}^*} \leq \overline{N}^* \tag{2.43}$$

We observe that this equation is identical to equation (2.12). Noting that $\rho_i^* = \gamma^* p_i \overline{x}_i$ and thus $\rho_{\max}^* = \gamma^* \max_{1 \leq j \leq M} p_j \overline{x}_j$ we have the following

Theorem 2.5

For the M/G/1 parallel network, the average number in system at maximum power satisfies

$$\sum_{i=1}^{M} \frac{p_i \overline{x}_i}{\max_{1 \leq j \leq M} p_j \overline{x}_j} \leq \overline{N}^* \tag{2.44}$$

This lower bound is given entirely in terms of the known quantities p_i and $\overline{x}_i$ $(= \overline{b}_i / C_i)$.

Using equation (2.41), we have found a lower bound on $\overline{N}^*$ for the M/G/1 parallel network, and we have also found that the unknown γ^* is a root of a certain polynomial and may be determined by any root finding procedure. However, for certain special cases of this network, we may exploit the equivalent equation (2.40) to obtain additional results about $\overline{N}^*$ (without explicitly solving for γ^*). This we now proceed to do.

2.4.1 The M/M/1 Parallel Network (Known Routing)

Assume all M channels are modeled as M/M/1 queueing systems (exponential message length with mean $\overline{b}_i = 1/\mu_i$). Then it is well-known that

$$\overline{N}_i^* = \frac{\rho_i^*}{1 - \rho_i^*}$$

for $1 \leq i \leq M$. Thus equation (2.40) becomes

$$\sum_{i=1}^{M} \overline{N}_i^* \left(1 - \overline{N}_i^*\right) = 0$$

which yields the following

Theorem 2.6

For the M/M/1 parallel network, the average number in system at maximum power satisfies

$$\overline{N}^* = \sum_{i=1}^{M} \overline{N}_i^* = \sum_{i=1}^{M} \left(\overline{N}_i^*\right)^2 \tag{2.45}$$

This expression is the same as Bharath-Kumar's equation (2.9) for the M/M/1 *series* network.

Proceeding in a manner identical to the proof of equation (2.8) from equation (2.9), we obtain the upper bound

$$\overline{N}^{*} \leq M$$

This is the same bound mentioned above for the M/M/1 tandem. Also, since $\bar{x}_i = 1/\mu_i C_i$ $(1 \leq i \leq M)$ for the M/M/1 parallel network, Theorem 2.5 (which holds for the more general M/G/1 parallel network) may be rewritten for M/M/1 as

$$\sum_{i=1}^{M} \frac{p_i/\mu_i C_i}{\max\limits_{1 \leq j \leq M} p_j/\mu_j C_j} \leq \overline{N}^{*}$$

or

$$\sum_{i=1}^{M} \frac{\min\limits_{1 \leq j \leq M} \mu_j C_j/p_j}{\mu_i C_i/p_i} \leq \overline{N}^{*}$$

If the (exponential) message lengths at all M channels are identical (i.e., we have $\mu_i \triangleq \mu$ for $1 \leq i \leq M$), the lower bound becomes

$$\sum_{i=1}^{M} \frac{\min\limits_{1 \leq j \leq M} C_j/p_j}{C_i/p_i} \leq \overline{N}^{*}$$

This lower bound for the M/M/1 parallel network was derived from results obtained for the M/G/1 parallel network; later in chapter 4 of this dissertation, we will obtain it as a particular case of a result obtained for M/M/1 networks with general topologies. We have shown the following (for equal message length distributions at all M channels)

Theorem 2.7

For the M/M/1 parallel network, the average number in system at maximum power satisfies

$$\sum_{i=1}^{M} \frac{\min\limits_{1 \leq j \leq M} C_j/p_j}{C_i/p_i} \leq \overline{N}^{*} \leq M \tag{2.46}$$

We observe that Theorem 2.7 is identical to Theorem 2.2 for the M/M/1 series network, since both theorems may be written in the form

$$\sum_{i=1}^{M} \frac{\rho_i^{*}}{\rho_{\max}^{*}} \leq \overline{N}^{*} \leq M \tag{2.47}$$

In chapter 4 we will show that these bounds hold for more general M/M/1 networks.

2.4.2 Equal Loads (Arbitrary Service Time Distributions)

We now assume that all M channels of our M/G/1 parallel network have equal loads; that is, $\rho_i = \rho_j \triangleq \rho$ for $1 \leq i,j \leq M$. This is similar to the series assumption of equal capacities, since in that case the assumption of equal capacities is equivalent to the assumption of equal loads (same input rate λ for all channels). Our point of departure is again equation (2.40). Since $\rho_i^* \triangleq \rho^*$ for $1 \leq i \leq M$, we may write that equation in the form

$$\frac{\rho^*}{1-\rho^*} \sum_{i=1}^{M} (1 - \overline{N}_i^*) = 0$$

Thus

$$\sum_{i=1}^{M} (1 - \overline{N}_i^*) = 0$$

or

$$\sum_{i=1}^{M} \overline{N}_i^* = M$$

Hence, regardless of the service time distributions at the individual nodes, the equal load assumption gives

Theorem 2.8

For the equal load M/G/1 parallel network, power is maximized when

$$\overline{N}^* = M \tag{2.48}$$

[Note: the special case of Theorem 2.8 when $M = 1$ gives Kleinrock's result that $\overline{N}^* = 1$ for M/G/1.] Here again we see invariance with respect to service time distribution. There is also an invariance with respect to load scaling similar to the invariance with respect to scaling of channel capacities for the M/D/1 series net.

To obtain other system parameters we first use equations (2.38) and (2.32) to write

$$\overline{N}^* = \sum_{i=1}^{M} \frac{\lambda_i^* W_i^*}{1-\rho_i^*} = \sum_{i=1}^{M} \left(\frac{\rho_i^*}{1-\rho_i^*}\right)^2 \left(\frac{1+\nu_i^2}{2}\right)$$

where, recall, ν_i is the coefficient of variation of service time for node i (i.e., ν_i is the ratio of the standard deviation to the mean service time). Since $\overline{N}^* = M$ and $\rho_i^* = \rho^*$ for all i, we have the quadratic equation

$$M(1-\rho^*)^2 = (\rho^*)^2 \sum_{i=1}^{M} \left(\frac{1+\nu_i^2}{2}\right)$$

Taking square roots, the channel utilization (equal to ρ^* for all M channels by assumption) is

$$\rho^* = \frac{\sqrt{M}}{\sqrt{M} + \sqrt{\sum_{i=1}^{M}(1+\nu_i^2)/2}} \tag{2.49}$$

To find the average number in system at the jth channel, we use equation (2.34) to write (recall that $\rho_j^* = \rho^*$)

$$\bar{N}_j^* = \rho^* + \frac{(\rho^*)^2}{1-\rho^*}\left(\frac{1+\nu_j^2}{2}\right) \tag{2.50}$$

or

$$\bar{N}_j^* = \rho^*\left[1 + \frac{\rho^*}{1-\rho^*}\left(\frac{1+\nu_j^2}{2}\right)\right]$$

Substituting equation (2.49) into this latter expression yields

$$\bar{N}_j^* = \frac{\sqrt{M}}{\sqrt{M} + \sqrt{\sum_{i=1}^{M}(1+\nu_i^2)/2}}\left[1 + \frac{\sqrt{M}}{\sqrt{\sum_{i=1}^{M}(1+\nu_i^2)/2}}\left(\frac{1+\nu_j^2}{2}\right)\right] \tag{2.51}$$

In general, the values of the average number at the *individual* nodes $\bar{N}_j^*$ will depend on the service time distribution (through ν_j^2). But amazingly, these quantities always add in such a way so as to force the total number $\bar{N}^* = M$.

In the case of equal coefficients of service time variation at the M nodes we can say more. Assume now that $\nu_i^2 \triangleq \nu^2$ for $1 \le i \le M$ (which is certainly true if all service time distributions are of the same type). Under this additional assumption equation (2.49) gives

$$\rho^* = \frac{1}{1+\sqrt{(1+\nu^2)/2}}$$

Also by equation (2.50) we clearly have

$$\bar{N}_i^* = \bar{N}_j^*$$

for $1 \le i,j \le M$, and so Theorem 2.8 yields

$$M = \bar{N}^* = \sum_{i=1}^{M} \bar{N}_i^* = M \cdot \bar{N}_i^*$$

Thus we have the following

Theorem 2.9

For the M/G/1 parallel network with equal loads and equal coefficients of service time variation

$$\bar{N}_i^* = 1 \qquad 1 \le i \le M \tag{2.52}$$

2.5 The Queueing System G/M/1

In this section the analysis of power is extended to the queueing system G/M/1. We find that the assumption of Poisson arrivals appears to be critical in the beautiful results for $\overline{N}^*$ derived above. However, we find a type of continuity property in the sense that, as long as the coefficient of variation ${\nu_a}^2$ of the interarrival process is close to 1 (i.e., Poisson arrivals), then $\overline{N}^*$ is approximately 1 as in M/G/1. But for large values of ${\nu_a}^2$, the value of $\overline{N}^*$ may become quite different than 1. In fact, a major theorem of this section on G/M/1 states that we can find G/M/1 systems so that at maximum power, $\overline{N}^*$ can be made arbitrarily close to 0. This is certainly different from Kleinrock's result for M/G/1.

Consider a G/M/1 queueing system with mean interarrival time $\overline{t}$ and mean (exponential) service time $\overline{x} = 1/\mu$. Note that $\lambda \triangleq 1/\overline{t}$ is the average arrival rate of customers (packets, messages) to the system, and thus the throughput γ is simply equal to λ. We begin our analysis of power for G/M/1 by recalling [Cohe69] that the equilibrium probabilities are

$$p_0 = 1 - \rho \qquad k = 0$$
$$p_k = \rho(1-\sigma)\sigma^{k-1} \qquad k \geq 1$$

where σ is the root between 0 and 1 satisfying the equation

$$\sigma = \hat{A}(\mu - \mu\sigma) \tag{2.53}$$

Here we break with the notation in [Klei75] and let $\hat{A}$ be the Laplace transform for the interarrival time density (the notation A^* of [Klei75] for the Laplace transform might be confusing, since we have previously introduced the convention that the exponent * indicates optimization with respect to power). We also recall that $\overline{N} = \rho/(1-\sigma)$ for G/M/1, and so $T = \overline{x}/(1-\sigma)$. In order to maximize power, we use equation (2.2) of Kleinrock

$$T^* = \gamma^* \left.\frac{dT}{d\gamma}\right|_{\gamma=\gamma^*} \tag{2.54}$$

which characterizes the maximum power point. However, to find the derivative $dT/d\gamma$, we note that T involves the root σ, which is a (possibly complicated) function of the throughput γ. For example, $\sigma = \rho = \gamma\overline{x}$ for M/M/1, while for E_2/M/1 (Erlang-2 interarrival process) it can be shown that

$$\sigma = \frac{1}{2}(1 + 4\rho - \sqrt{1+8\rho})$$

Thus $dT/d\gamma$ will involve the derivative $d\sigma/d\gamma$. Let us assume this latter derivative exists (in all the examples of G/M/1 systems that we analyze, this will be the case). Since $T = \overline{x}/(1-\sigma)$, then

$$\frac{dT}{d\gamma} = \frac{\overline{x}}{(1-\sigma)^2} \cdot \frac{d\sigma}{d\gamma} = \frac{T}{1-\sigma} \cdot \frac{d\sigma}{d\gamma} \tag{2.55}$$

Substituting equation (2.55) into equation (2.54) yields

$$T^* = \frac{\gamma^* T^*}{1-\sigma^*} \cdot \left.\frac{d\sigma}{d\gamma}\right|_{\gamma=\gamma^*}$$

Using Little's result we have

$$T^*(1-\sigma^*) = \overline{N}^* \cdot \left.\frac{d\sigma}{d\gamma}\right|_{\gamma=\gamma^*}$$

or

$$\bar{x} = \overline{N}^* \cdot \left.\frac{d\sigma}{d\gamma}\right|_{\gamma=\gamma^*} \tag{2.56}$$

This equation relates various parameters at the maximum power point for the queue G/M/1. Note that for M/M/1 (Poisson arrivals) we have $\sigma = \rho = \gamma\bar{x}$ and so $d\sigma/d\gamma = \bar{x}$. Thus equation (2.56) reduces to the result of Kleinrock that $\overline{N}^* = 1$ for M/M/1. We now use equation (2.56) to analyze various interarrival processes.

2.5.1 Erlangian Input

The first G/M/1 system we analyze has a k-stage Erlangian interarrival time process. Here $k\lambda$ is the (Poisson) arrival rate to each of the k stages, so that

$$\bar{t} = k \cdot \frac{1}{k\lambda} = \frac{1}{\lambda}$$

Thus the arrival rate to the system is λ and the throughput is $\gamma = \lambda$. The utilization is $\rho = \lambda/\mu = \gamma/\mu$. The variance of the interarrival time distribution is

$$VAR = k \cdot \frac{1}{(k\lambda)^2} = \frac{1}{k\lambda^2}$$

The coefficient of variation for the arrival process is

$$\nu_a^2 = \frac{VAR}{(\bar{t})^2} = \frac{1}{k}$$

and is therefore less than or equal to 1. The Laplace transform $\hat{A}$ is given by

$$\hat{A}(s) = \left(\frac{k\lambda}{s+k\lambda}\right)^k$$

and thus σ is found by solving the equation

$$\sigma = \hat{A}(\mu - \mu\sigma) = \left(\frac{k\lambda}{\mu - \mu\sigma + k\lambda}\right)^k$$

Using $\rho = \lambda/\mu$ this becomes

$$\sigma = \left(\frac{k\rho}{1-\sigma+k\rho}\right)^k \tag{2.57}$$

We shall now find $d\sigma/d\gamma$. Differentiating both sides of equation (2.57) with respect to γ yields

$$\frac{d\sigma}{d\gamma} = k\left(\frac{k\rho}{1-\sigma+k\rho}\right)^{k-1} \frac{(1-\sigma+k\rho)k\bar{x} - k\rho\left(-\frac{d\sigma}{d\gamma}+k\bar{x}\right)}{(1-\sigma+k\rho)^2}$$

Multiplying both sides of this equation by ρ and simplifying gives

$$\rho\frac{d\sigma}{d\gamma} = \left(\frac{k\rho}{1-\sigma+k\rho}\right)^{k} \frac{(1-\sigma)k\bar{x} + k\rho\frac{d\sigma}{d\gamma}}{1-\sigma+k\rho}$$

Therefore, using equation (2.57), we have

$$\frac{d\sigma}{d\gamma}[\rho(1-\sigma+k\rho)] = \sigma\left[(1-\sigma)k\bar{x} + k\rho\frac{d\sigma}{d\gamma}\right]$$

or

$$\frac{d\sigma}{d\gamma}[\rho(1-\sigma+k\rho) - k\rho\sigma] = \sigma(1-\sigma)k\bar{x}$$

We finally have

$$\frac{d\sigma}{d\gamma} = \frac{k\bar{x}\sigma(1-\sigma)}{\rho(1-\sigma+k\rho-k\sigma)} \tag{2.58}$$

Note that, for $k=1$, we have an M/M/1 system, so that $\sigma=\rho$. In this case equation (2.58) reduces to $d\sigma/d\gamma = \bar{x}$ as expected.

We now use this expression for $d\sigma/d\gamma$ in equation (2.56) to find the optimal power point. We have

$$\bar{x} = \frac{\rho^*}{1-\sigma^*} \cdot \frac{k\bar{x}\sigma^*(1-\sigma^*)}{\rho^*(1-\sigma^*+k\rho^*-k\sigma^*)}$$

which simplifies to

$$1 = \frac{k\sigma^*}{1-\sigma^*+k\rho^*-k\sigma^*}$$

This yields the equation

$$1-\sigma^*+k\rho^*-k\sigma^* = k\sigma^*$$

or

$$1-\sigma^*+k\rho^* = 2k\sigma^*$$

Thus we have

$$\sigma^* = \frac{1+k\rho^*}{2k+1} \tag{2.59}$$

and also

$$\rho^* = \frac{2k\sigma^* - (1-\sigma^*)}{k} \tag{2.60}$$

Using the above expression for ρ^* in equation (2.57), we see that σ^* must satisfy

$$\sigma^* = \left[\frac{2k\sigma^* - (1-\sigma^*)}{2k\sigma^*}\right]^k \tag{2.61}$$

which is equivalent to

$$\sigma^* = \left[1 - \frac{(1-\sigma^*)/2\sigma^*}{k}\right]^k \tag{2.62}$$

For any value of k, equation (2.61) (or (2.62)) can be solved numerically for σ^*. Then values of other system parameters at the optimal point (such as ρ^* and $\overline{N}^*$) can be found.

As an example, we now analyze the system $E_2/M/1$ (the case $k = 2$). For this system, equation (2.61) is

$$\sigma^* = \left[\frac{4\sigma^* - (1-\sigma^*)}{4\sigma^*}\right]^2$$

which yields

$$16(\sigma^*)^3 = (5\sigma^* - 1)^2 = 25(\sigma^*)^2 - 10\sigma^* + 1$$

or

$$16(\sigma^*)^3 - 25(\sigma^*)^2 + 10\sigma^* - 1 = 0$$

Since $\sigma^* = 1$ is always a root, we factor the above equation and find

$$(\sigma^* - 1)[16(\sigma^*)^2 - 9\sigma^* + 1] = 0$$

Thus σ^* must be a root of the equation

$$16(\sigma^*)^2 - 9\sigma^* + 1 = 0$$

The two roots of this quadratic are

$$\sigma^* = \frac{9 \pm \sqrt{17}}{32}$$

Now from equation (2.60) we know that ρ^* satisfies

$$\rho^* = \frac{4\sigma^* - (1-\sigma^*)}{2} = \frac{5\sigma^* - 1}{2}$$

Hence the above two possible values of σ^* yield the corresponding values

$$\rho^* = \frac{13 \pm 5\sqrt{17}}{64}$$

Since the negative square root gives $\rho^* < 0$, we must choose the positive square root. Therefore

$$\sigma^* = \frac{9+\sqrt{17}}{32}$$

and

$$\rho^* = \frac{13+5\sqrt{17}}{64}$$

We also have

$$\overline{N}^* = \frac{\rho^*}{1-\sigma^*} = \frac{13+5\sqrt{17}}{46-2\sqrt{17}}$$

Thus, for the system $E_2/M/1$, we have

$$\sigma^* \cong .410097, \quad \rho^* \cong .525243, \quad \overline{N}^* \cong .890388 \tag{2.63}$$

We note that the relationship $\overline{N}^* = 1$ (which holds for M/G/1) is not true for this $E_2/M/1$ system.

We next consider the system D/M/1 (i.e., the case of constant interarrival times and exponential service times). For this system we have $\nu_a^2 = 0$. Since $D/M/1 = \lim_{k\to\infty} E_k/M/1$, we may analyze this queue by using our previous results concerning power for the Erlangian systems. Thus, as $k \to \infty$, equation (2.60) provides the interesting result

$$\rho^* = 2\sigma^* \tag{2.64}$$

which holds for D/M/1 at the optimal power point. Also equation (2.62) becomes, in the limit,

$$\sigma^* = e^{-[(1-\sigma^*)/2\sigma^*]} \tag{2.65}$$

Solving this equation for σ^* (numerically) and using the result to calculate ρ^* and $\overline{N}^*$ we obtain for the D/M/1 queue

$$\sigma^* \cong .284668, \quad \rho^* \cong .569336, \quad \overline{N}^* \cong .795905 \tag{2.66}$$

These equations may also be obtained directly without resorting to the $E_k/M/1$ results. Since $\hat{A}(s) = e^{-s\overline{x}}$ for D/M/1, equation (2.53) gives

$$\sigma = e^{-[(1-\sigma)/\rho]} = e^{-1/\overline{N}}$$

Differentiating this equation with respect to γ, we find (after considerable computation)

$$\frac{d\sigma}{d\gamma} = \frac{\sigma(1-\sigma)\overline{x}}{\rho(\rho-\sigma)}$$

Equation (2.56) then yields $\rho^* = 2\sigma^*$ as before.

We now list values of parameters for the system $E_k/M/1$ at maximum power and compare them to the corresponding $M/E_k/1$ system. Recall that, for M/G/1,

$$\overline{N}^* = 1$$

and

$$\rho^* = \frac{1}{1+\sqrt{(1+\nu_b{}^2)/2}}$$

where $\nu_b{}^2$ is the squared coefficient of variation for the (general) service time distribution. Thus for $M/E_k/1$ (with $\nu_b{}^2 = 1/k$) we have

$$\rho^* = \frac{1}{1+\sqrt{(k+1)/(2k)}}$$

In this following table we note that, even for constant interarrival time ($k = \infty$), $\overline{N}^*$ remains fairly close to 1 (at about .8).

	$M/E_k/1$		$E_k/M/1$		
k	ρ^*	$\overline{N}^*$	ρ^*	σ^*	$\overline{N}^*$
1	1/2	1	1/2	1/2	1
2	.535898	1	.525243	.410097	.890388
3	.550510	1	.537006	.373002	.856472
4	.558482	1	.543813	.352806	.840263
5	.563508	1	.548252	.340115	.830829
10	.574178	1	.558058	.313361	.812739
25	.581020	1	.564628	.296386	.802468
50	.583382	1	.566947	.290568	.799156
100	.584579	1	.568133	.287628	.797523
1000	.585665	1	.569215	.284965	.796066
∞	.585786	1	.569336	.284668	.795905

Table 2.3

For these $E_k/M/1$ systems, the squared coefficient of variation of interarrival time is between 0 and 1 ($\nu_a^2 = 1/k$). We may consider more general interarrival processes which are sums of exponential phases with unequal parameters λ_i (the so-called *generalized Erlangian* process). In this case, the interarrival time transform is

$$\hat{A}(s) = \prod_{i=1}^{k} \frac{\lambda_i}{s+\lambda_i}$$

The mean interarrival time is

$$\bar{t} = \sum_{i=1}^{k} \frac{1}{\lambda_i}$$

while the variance of interarrival time is

$$VAR = \sum_{i=1}^{k} \frac{1}{(\lambda_i)^2}$$

The squared coefficient of variation of interarrival time is

$$\nu_a^2 = \frac{VAR}{(\bar{t})^2} = \frac{\sum_{i=1}^{k} \left(\frac{1}{\lambda_i}\right)^2}{\left(\sum_{i=1}^{k} \frac{1}{\lambda_i}\right)^2}$$

and thus

$$\nu_a^2 = \frac{\sum_{i=1}^{k} \left(\frac{1}{\lambda_i}\right)^2}{\sum_{i=1}^{k} \left(\frac{1}{\lambda_i}\right)^2 + \sum\sum_{j \neq k} \frac{1}{\lambda_j} \cdot \frac{1}{\lambda_k}} \leq 1$$

Once again our $G/M/1$ systems satisfy $\nu_a^2 \leq 1$.

We now specifically consider such systems for which $\lambda_i = \lambda/\beta_i$, where β_i is a fixed constant satisfying $0 < \beta_i < 1$ and where $\sum_{i=1}^{M} \beta_i = 1$. Note that the $E_k/M/1$ system is an example with $\beta_i = 1/k$ for $1 \leq i \leq k$. Our Erlangian analysis may be extended to these newly defined systems as follows. We first observe that the mean interarrival time is

$$\bar{t} = \sum_{i=1}^{k} \frac{1}{\lambda_i} = \sum_{i=1}^{k} \frac{\beta_i}{\lambda} = \frac{1}{\lambda}$$

Thus the arrival rate is again λ and so the throughput is $\gamma = \lambda$. The utilization is simply $\rho = \lambda/\mu = \gamma/\mu$. The variable σ is the root between 0 and 1 of the equation

$$\sigma = \prod_{i=1}^{k} \frac{\lambda/\beta_i}{\mu - \mu\sigma + \lambda/\beta_i}$$

which is equivalent to

$$\sigma = \prod_{i=1}^{k} \frac{\rho/\beta_i}{1 - \sigma + \rho/\beta_i} \tag{2.67}$$

Since the β_i are fixed constants, it is easy to differentiate equation (2.67) with respect to γ to yield $d\sigma/d\gamma$. Then equation (2.56) may be used to solve for ρ^* and σ^* as was done for the Erlangian case.

The coefficient of variation of interarrival time was less than or equal to one for all of the above G/M/1 systems. If we now look at systems with $\nu_a^2 \geq 1$ we find that the values of various parameters at maximum power become quite different from those for the Poisson case. We will next study systems with hyperexponential interarrival time distributions (these systems satisfy $\nu_a^2 \geq 1$).

2.5.2 Hyperexponential Input

We first consider a general hyperexponential interarrival process consisting of R parallel stages, with the probability of choosing stage i to be α_i, and stage i is exponential with parameter λ_i. We assume that $R > 1$, since we have an M/M/1 system for $R = 1$. We also assume that $0 < \alpha_i < 1$ for all $1 \leq i \leq R$ and that $\sum_{i=1}^{R} \alpha_i = 1$ (see Figure 2.7).

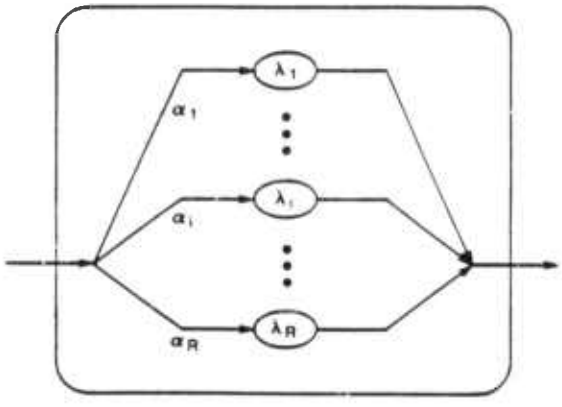

Figure 2.7 The Hyperexponential Interarrival Process

The Laplace transform for this interarrival time process satisfies

$$\hat{A}(s) = \sum_{i=1}^{R} (\alpha_i) \frac{\lambda_i}{s + \lambda_i}$$

and the first two moments are therefore

$$\bar{t} = \sum_{i=1}^{R} (\alpha_i) \frac{1}{\lambda_i}$$

and

$$\overline{t^2} = \sum_{i=1}^{R} (\alpha_i) \frac{2}{(\lambda_i)^2}$$

Thus ν_e^2, the squared coefficient of variation of the interarrival time distribution, satisfies

$$\nu_e^2 = \frac{\overline{t^2}}{(\bar{t})^2} - 1 = 2 \frac{\sum_{i=1}^{R} \frac{\alpha_i}{\lambda_i^2}}{\left(\sum_{i=1}^{R} \frac{\alpha_i}{\lambda_i}\right)^2} - 1$$

Using the Cauchy-Schwarz inequality

$$\left(\sum_i x_i y_i\right)^2 \leq \left(\sum_i x_i^2\right)\left(\sum_i y_i^2\right)$$

with

$$x_i = \frac{\sqrt{\alpha_i}}{\lambda_i}$$

and

$$y_i = \sqrt{\alpha_i}$$

as in [Klei75], we find that $\nu_e^2 \geq 1$.

We now restrict our analysis to the following $H_R/M/1$ systems. We consider systems for which $\lambda_i = \lambda(\alpha_i/\beta_i)$ where β_i is a fixed constant with $0 < \beta_i < 1$ and where $\sum_{i=1}^{R} \beta_i = 1$. Note that the mean interarrival time is

$$\bar{t} = \sum_{i=1}^{R} (\alpha_i) \frac{1}{\lambda_i} = \sum_{i=1}^{R} (\alpha_i) \frac{\beta_i}{\lambda \alpha_i} = \frac{1}{\lambda}$$

Thus λ is the arrival rate to the system, and so the throughput is $\gamma = \lambda$. The utilization is $\rho = \lambda/\mu = \gamma/\mu$. To find the root σ we solve equation (2.53). This becomes

$$\sigma = \sum_{i=1}^{R} (\alpha_i) \frac{\lambda(\alpha_i/\beta_i)}{\mu - \mu\sigma + \lambda(\alpha_i/\beta_i)}$$

or, writing the above in terms of ρ,

$$\sigma = \sum_{i=1}^{R} \frac{\rho(\alpha_i^2/\beta_i)}{1 - \sigma + \rho(\alpha_i/\beta_i)} \tag{2.68}$$

For every value of the input rate λ (or the utilization ρ), the corresponding value of σ is determined from equation (2.68). In particular, the optimal value σ^* is given by that equation if the optimal efficiency ρ^* is known.

We now use equation (2.56) to find another expression relating ρ^* and σ^*. We first determine $d\sigma/d\gamma$. Since the α_i and the β_i are fixed constants, we may easily differentiate both sides of equation (2.68) with respect to γ to obtain

$$\frac{d\sigma}{d\gamma} = \sum_{i=1}^{R} \frac{[1-\sigma+\rho(\alpha_i/\beta_i)][\bar{x}(\alpha_i^2/\beta_i)] - [\rho(\alpha_i^2/\beta_i)]\left[-\frac{d\sigma}{d\gamma} + \bar{x}(\alpha_i/\beta_i)\right]}{[1-\sigma+\rho(\alpha_i/\beta_i)]^2}$$

or

$$\frac{d\sigma}{d\gamma} = \sum_{i=1}^{R} \frac{(1-\sigma)\bar{x}(\alpha_i^2/\beta_i) + \rho(\alpha_i^2/\beta_i)\frac{d\sigma}{d\gamma}}{[1-\sigma+\rho(\alpha_i/\beta_i)]^2}$$

This equation becomes

$$\frac{d\sigma}{d\gamma}\left[1 - \sum_{i=1}^{R} \frac{\rho(\alpha_i^2/\beta_i)}{[1-\sigma+\rho(\alpha_i/\beta_i)]^2}\right] = \sum_{i=1}^{R} \frac{(1-\sigma)\bar{x}(\alpha_i^2/\beta_i)}{[1-\sigma+\rho(\alpha_i/\beta_i)]^2}$$

or

$$\frac{d\sigma}{d\gamma} = \frac{\sum_{i=1}^{R} \frac{(1-\sigma)\bar{x}(\alpha_i^2/\beta_i)}{[1-\sigma+\rho(\alpha_i/\beta_i)]^2}}{1 - \sum_{i=1}^{R} \frac{\rho(\alpha_i^2/\beta_i)}{[1-\sigma+\rho(\alpha_i/\beta_i)]^2}}$$

We finally have

$$\frac{d\sigma}{d\gamma} = \frac{(1-\sigma)\bar{x}\sum_{i=1}^{R} \frac{\alpha_i^2/\beta_i}{[1-\sigma+\rho(\alpha_i/\beta_i)]^2}}{1 - \rho\sum_{i=1}^{R} \frac{\alpha_i^2/\beta_i}{[1-\sigma+\rho(\alpha_i/\beta_i)]^2}} \tag{2.69}$$

We now use this expression for $d\sigma/d\gamma$ in equation (2.56) to find the optimal power point. We have

$$\bar{x} = \frac{\rho^*}{1-\sigma^*} \cdot \frac{(1-\sigma^*)\bar{x}\sum_{i=1}^{R} \frac{\alpha_i^2/\beta_i}{[1-\sigma^*+\rho^*(\alpha_i/\beta_i)]^2}}{1 - \rho^*\sum_{i=1}^{R} \frac{\alpha_i^2/\beta_i}{[1-\sigma^*+\rho^*(\alpha_i/\beta_i)]^2}}$$

or

$$1 - \rho^*\sum_{i=1}^{R} \frac{\alpha_i^2/\beta_i}{[1-\sigma^*+\rho^*(\alpha_i/\beta_i)]^2} = \rho^*\sum_{i=1}^{R} \frac{\alpha_i^2/\beta_i}{[1-\sigma^*+\rho^*(\alpha_i/\beta_i)]^2} \tag{2.70}$$

Further simplifying equations (2.68) and (2.70), we have two equations in the two unknowns ρ^* and σ^*, namely

$$\sigma^* = \rho^* \sum_{i=1}^{R} \frac{(\alpha_i^2/\beta_i)}{1-\sigma^*+\rho^*(\alpha_i/\beta_i)} \tag{2.71}$$

and

$$1 = 2\rho^* \sum_{i=1}^{R} \frac{\alpha_i^2/\beta_i}{[1-\sigma^*+\rho^*(\alpha_i/\beta_i)]^2} \tag{2.72}$$

These two equations can be solved numerically for any R and any values of α_i and β_i which satisfy the conditions mentioned above. This yields σ^* and ρ^* and thus $\bar{N}^* = \rho^*/(1-\sigma^*)$.

2.5.3 Keep The Pipe Full Counterexample

Previous results show that, although Kleinrock's "keep the pipe full" result for M/G/1 does not extend to G/M/1, the average number in system at optimal power $\bar{N}^*$ is fairly close to 1 for systems with Erlangian interarrival time distributions. However, it was also noted that the value of ν_e^2 for these systems was between zero and one. We now consider hyperexponential distributions with an arbitrarily large coefficient of variation for the interarrival process. We choose certain α_i and β_i and solve for $\bar{N}^*$ to show that we can make $\bar{N}^*$ as small as desired ($\bar{N}^*$ positive). This gives the theorem promised above.

The interarrival time process will consist of two parallel phases (an $H_2/M/1$ queueing system). We set $\alpha_1 = \delta$ and $\alpha_2 = 1-\delta$, where δ is a real number which satisfies $0 < \delta < 1$. We next set $\beta_1 = (K-1)/K$ and $\beta_2 = 1/K$, where K is a positive integer greater than 1 (i.e., $K \geq 2$). Note that $\alpha_1 + \alpha_2 = 1 = \beta_1 + \beta_2$. We have

$$\bar{t} = \frac{1}{\lambda}$$

Thus the arrival rate is λ, the throughput is $\gamma = \lambda$ and the utilization is $\rho = \lambda/\mu = \gamma/\mu$. Also

$$\nu_e^2 = 2\left[\frac{(K-1)^2}{K^2\delta} + \frac{1}{K^2(1-\delta)}\right] - 1$$

We observe that ν_e^2 becomes arbitrarily large as $\delta \to 0$ (for fixed K). The interarrival process is illustrated in Figure 2.8 below.

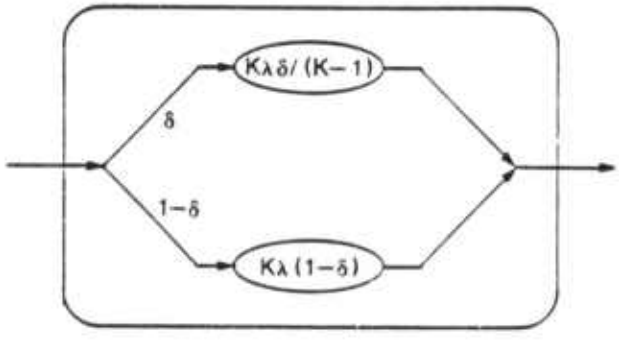

Figure 2.8 An Interarrival Process for Bursts of Messages

With low probability, the interarrival time between messages will be large, while the interarrival time will be small with high probability. Such an arrival process can be used to model a very bursty source of messages (such as a terminal). In such a situation, the think time between bursts of messages may be large compared to the time between messages from the same burst. Thus the interarrival process will approximate the case of bulk arrivals.

We first examine the behavior of σ for a fixed value of ρ when the parameter δ becomes small. Under the above choice of the α_i and β_i, equation (2.68) yields

$$\sigma = \frac{\rho\delta^2 K/(K-1)}{1-\sigma+\rho\delta K/(K-1)} + \frac{\rho(1-\delta)^2 K}{1-\sigma+\rho(1-\delta)K}$$

or

$$\sigma = K\rho\left[\frac{\delta^2/(K-1)}{1-\sigma+K\rho\delta/(K-1)} + \frac{(1-\delta)^2}{1-\sigma+K\rho(1-\delta)}\right]$$

This equation becomes

$$\sigma[1-\sigma+K\rho\delta/(K-1)][1-\sigma+K\rho(1-\delta)] =$$

$$[K\rho\delta^2/(K-1)][1-\sigma+K\rho(1-\delta)] + [K\rho(1-\delta)^2][1-\sigma+K\rho\delta/(K-1)]$$

which is equivalent to

$$\sigma(1-\sigma)^2 + \sigma(1-\sigma)[K\rho\delta/(K-1) + K\rho(1-\delta)] + \sigma[(K\rho)^2\delta(1-\delta)/(K-1)] =$$

$$(1-\sigma)K\rho\delta^2/(K-1) + (1-\sigma)K\rho(1-\delta)^2 + (K\rho)^2\delta^2(1-\delta)/(K-1) + (K\rho)^2\delta(1-\delta)^2/(K-1)$$

The right-hand side of the above equation may be simplified by noting that

$$(K\rho)^2\delta^2(1-\delta)/(K-1) + (K\rho)^2\delta(1-\delta)^2/(K-1) = (K\rho)^2\delta(1-\delta)/(K-1)$$

This yields

$$\sigma(1-\sigma)^2 + \sigma(1-\sigma)[K\rho\delta/(K-1) + K\rho(1-\delta)] =$$

$$(1-\sigma)K\rho\delta^2/(K-1) + (1-\sigma)K\rho(1-\delta)^2 + (1-\sigma)(K\rho)^2\delta(1-\delta)/(K-1)$$

It is no surprise that $1-\sigma$ is a common factor, because the above equation was derived from equation (2.68) which is a special instance of the general G/M/1 equation $\sigma = \hat{A}(\mu - \mu\sigma)$. Dividing by $1-\sigma$, we have

$$\sigma(1-\sigma) + \sigma[K\rho\delta/(K-1) + K\rho(1-\delta)] = K\rho\delta^2/(K-1) + K\rho(1-\delta)^2 + (K\rho)^2\delta(1-\delta)/(K-1)$$

or

$$\sigma(1-\sigma) = K\rho(1-\delta)(1-\sigma-\delta) + K\rho\delta/(K-1)[\delta + K\rho(1-\delta) - \sigma]$$

This is equivalent to

$$\sigma(1-\sigma) = K\rho(1-\sigma) - K\rho\delta(1-\sigma) - K\rho\delta(1-\delta) + K\rho\delta/(K-1)[1-\sigma-(1-\delta)+K\rho(1-\delta)]$$

or

$$\sigma(1-\sigma) = K\rho(1-\sigma) - K\rho\delta(1-\sigma)\left(1-\frac{1}{K-1}\right) - K\rho\delta(1-\delta)\left(1+\frac{1}{K-1}-\frac{K\rho}{K-1}\right)$$

We finally have

$$\sigma(1-\sigma) = K\rho(1-\sigma) - K\rho\delta(1-\sigma)\frac{K-2}{K-1} - K\rho\delta(1-\delta)\frac{K(1-\rho)}{K-1}$$

or

$$K\rho(1-\sigma) = \sigma(1-\sigma) + \frac{K\rho\delta}{K-1}[(K-2)(1-\sigma)+K(1-\delta)(1-\rho)]$$

This may be written in the form

$$K\rho(1-\sigma) = \sigma(1-\sigma) + f(\delta) \tag{2.73}$$

where

$$f(\delta) \triangleq \frac{K\rho\delta}{K-1}[(K-2)(1-\sigma)+K(1-\delta)(1-\rho)] \tag{2.74}$$

Note that equation (2.73) is equivalent to

$$(K\rho-\sigma)(1-\sigma) = f(\delta) \tag{2.75}$$

Since $K \geq 2$, we observe that

$$0 < f(\delta) < \frac{K\delta}{K-1}[(K-2)+K] = 2K\delta$$

for $0<\rho<1$ and $0<\delta<1$. Thus we have $\lim_{\delta\to 0} f(\delta) = 0$ for any (all) ρ between 0 and 1. We also have

$$0 < (K\rho-\sigma)(1-\sigma) < 2K\delta \tag{2.76}$$

from equation (2.75) for this range of ρ and δ. Using the left-hand inequality (and the fact that σ lies between 0 and 1), we obtain

$$0 < \sigma < \min(K\rho, 1) \tag{2.77}$$

We consider a fixed $K \geq 2$ and a fixed δ satisfying $0<\delta<1$. Note that equation (2.76) is valid for all $0<\rho<1$. We now examine the behavior of σ for the two ranges $0<\rho\leq 1/K$ and $1/K \leq \rho < 1$. For $\rho = 1/K$, equation (2.76) yields

$$0 < (1-\sigma)^2 < 2K\delta$$

or

$$0 < 1-\sigma < \sqrt{2K\delta}$$

Since σ is an increasing function of the input rate (σ represents the probability that a customer arrives to find a non-empty system [Klei75]), we have

$$0 < 1-\sigma < \sqrt{2K\delta} \tag{2.78}$$

for $1/K \leq \rho < 1$. Now consider the range $0 < \rho \leq 1/K$. For those values of ρ, $\sigma < K\rho$ by equation (2.77). We now use $K\rho \leq 1$ and equation (2.76) to obtain

$$0 < (K\rho-\sigma)^2 \leq (K\rho-\sigma)(1-\sigma) < 2K\delta$$

Thus

$$0 < K\rho-\sigma < \sqrt{2K\delta} \tag{2.79}$$

for $0 < \rho \leq 1/K$. From equations (2.78) and (2.79), the behavior of σ with respect to ρ becomes evident for small δ. We have $\sigma \cong K\rho$ ($\sigma < K\rho$) for $0 < \rho \leq 1/K$, and $\sigma \cong 1$ ($\sigma < 1$) for $1/K \leq \rho < 1$.

The power function satisfies

$$P = \frac{\gamma}{T} = \frac{\gamma^2}{\gamma T} = \frac{1}{(\bar{x})^2} \cdot \frac{\rho^2}{\bar{N}}$$

and so, since $\bar{N} = \rho/(1-\sigma)$ for G/M/1,

$$P = \frac{\rho(1-\sigma)}{(\bar{x})^2} \tag{2.80}$$

To find the optimal power point, we need to maximize P. For $1/K \leq \rho < 1$, equation (2.78) yields

$$(\bar{x})^2 P = \rho(1-\sigma) < 1-\sigma < \sqrt{2K\delta}$$

Now consider the particular point $\rho = 1/2K$. For this value of ρ, equation (2.77) yields $\sigma < 1/2$. Thus we have $1-\sigma > 1/2$, and so

$$(\bar{x})^2 P = \rho(1-\sigma) > \frac{1}{4K} > \sqrt{2K\delta}$$

for small δ. Therefore, the optimal power point cannot occur in the range $1/K \leq \rho < 1$, and so it must occur for some ρ satisfying $0 < \rho < 1/K$. For this range, we recall that $\sigma \cong K\rho$, and therefore

$$P \cong \frac{\rho(1-K\rho)}{(\bar{x})^2} = \frac{\rho - K\rho^2}{(\bar{x})^2}$$

Differentiating the right-hand expression, we find that the value of ρ which optimizes power is

$$\rho^* \cong \frac{1}{2K} \tag{2.81}$$

Since $\sigma \cong K\rho$ in this range, we find

$$\sigma^* \cong \frac{1}{2} \tag{2.82}$$

Since $\bar{N} = \rho/(1-\sigma)$ for G/M/1, we finally have

$$\bar{N}^* \cong \frac{1}{K} \tag{2.83}$$

Thus for any fixed K, we can find a $\delta(K) > 0$ so that $\bar{N}^* \cong 1/K$. We also have

$$P = \frac{\rho(1-\sigma)}{(\bar{x})^2} \cong \frac{1}{4K(\bar{x})^2}$$

We now can show the following

Theorem 2.10

Given $\epsilon > 0$, there is a G/M/1 system (in fact an H_2/M/1 system) such that at maximum power

$$0 < \bar{N}^* < \epsilon \tag{2.84}$$

To prove this we first choose K so large that $1/K < \epsilon$. Then we choose $\delta(K) > 0$ so that we may construct an H_2/M/1 system as above with $\bar{N}^* \cong 1/K$. That is, we may choose K and then $\delta(K)$ so that $0 < \bar{N}^* < \epsilon$.

In this chapter, we have seen that the average number in system for certain simple networks operating at maximum power is a quantity that exhibits invariances in several ways and is numerically quite easy to evaluate. Not only were these networks topologically simple, the power problems which were studied turned out to be mathematically simple. When more complicated network problems are considered, we will find that maximizing power may be difficult. We will extend our analysis of power to such problems in the next chapter and discover several undesirable characteristics of the optimal operating point. However, in later chapters, some of the beautiful results obtained above for these simple networks will be extended to more complex power problem formulations.

CHAPTER 3
Power of the M/M/1 Parallel Network

In the previous chapter, several simple networks were analyzed in terms of the maximization of power. These networks were not only topologically simple, they were also "simple" from an optimization point of view. That is, they all yielded single-variable optimization problems, and thus the equation

$$T = \gamma \frac{dT}{d\gamma}$$

could be used to analyze the optimal solution which maximized power. One of the networks considered was an M/G/1 parallel net where the fraction of traffic p_i on channel i was known. The resulting optimization problem involved only a single variable, the total input γ to the parallel system. The individual traffic λ_i on channel i was simply $\lambda_i = p_i \gamma$. In this chapter we examine another optimization problem involving the maximization of power for a parallel net with M channels. However, unlike the single variable parallel network problem of chapter 2, this new formulation is a multi-variable optimization problem, whose decision variables are the M individual channel flows λ_i, which are to be independently adjusted.

3.1 Description of the Optimization Problem

Consider now a parallel net with M channels which, unlike the above-mentioned system, has no restriction on the routing, and let us optimize it with respect to power. We can use this parallel network to represent either a single user with multiple paths for his packets or multiple users, each with a single path. In the first interpretation, the net is used to model a single user of a computer network (i.e., a single source-destination pair) with M alternate paths (possibly virtual circuits) for the user's packets to travel through the net. Unlike the previous models, the routing for this net is not known and is a system variable which must be optimized. This network may also represent multiple users (source-destination pairs) in a computer network. The flows on the channels would then represent the amount of traffic input to the system by each user. Note that in this latter interpretation, the users are independent (their packets don't interfere with one another). The corresponding system parameters of the users only interact in the calculation of the power function P. Finally, we note that analysis of the parallel net is important, since we may use it along with the series net explored above as building blocks for a general network.

Let us now state the optimization problem we wish to solve. We are given M parallel channels with capacities $C_1, \ldots, C_M$, and we assume that the message length distribution is exponential with mean $\bar{b} = 1/\mu$. We wish to find inputs $\lambda_1, \ldots, \lambda_M$ which maximize the power of the system (see Figure 3.1).

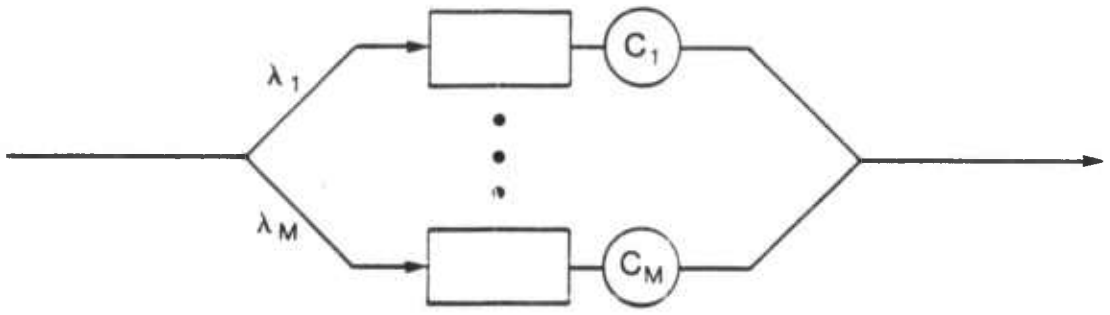

Figure 3.1 The M/M/1 Parallel Network

Note that the problem being considered, unlike the single-variable problems studied previously, involves optimizing the M variables $\lambda_1, \ldots, \lambda_M$. To obtain the objective function of our optimization problem (the power function P) in terms of the λ_i, we first note that the system throughput is given by

$$\gamma = \lambda_1 + \cdots + \lambda_M$$

We may express the mean delay in the system T as

$$T = \sum_{i=1}^{M} \frac{\lambda_i}{\gamma} T_i$$

which becomes, using the value for γ given above,

$$T = \frac{\lambda_1}{\lambda_1 + \cdots + \lambda_M} T_1 + \cdots + \frac{\lambda_M}{\lambda_1 + \cdots + \lambda_M} T_M$$

Thus

$$T = \frac{\sum_{i=1}^{M} \lambda_i T_i}{\sum_{i=1}^{M} \lambda_i}$$

We have expressed both the throughput γ and the delay T in terms of the decision variables $\lambda_1, \ldots, \lambda_M$. Using these expressions, we find that the power function $P = \gamma/T$ is

$$P = \frac{\left(\sum_{i=1}^{M} \lambda_i\right)^2}{\sum_{i=1}^{M} \lambda_i T_i} \tag{3.1}$$

We now extend the function P of equation (3.1) in a continuous manner so that the feasible region of our optimization problem is compact (closed and bounded) by considering the two (extreme) cases of zero throughput and infinite delay. We know that if some $\lambda_i \to \mu C_i$, the (M/M/1) delay $T_i = 1/(\mu C_i - \lambda_i) \to \infty$, and so from equation (3.1) we must have $P \to 0$. We now consider the case when $\lambda_i \to 0$ for all i (the case of zero throughput). Since

$$T_i \geq \bar{x}_i \geq \min_{1 \leq j \leq M} \bar{x}_j = \frac{1}{\mu C_{\max}} > 0$$

for $1 \leq i \leq M$, equation (3.1) yields

$$0 \leq P \leq \frac{\left(\sum_{i=1}^{M} \lambda_i\right)^2}{\sum_{i=1}^{M} \frac{\lambda_i}{\mu C_{\max}}} = \mu C_{\max} \sum_{i=1}^{M} \lambda_i$$

and therefore $P \to 0$ when all λ_i approach zero. We may thus extend equation (3.1) for P to a continuous function which includes these limiting cases when the network has zero power. The feasible region is then considered to be the set of all vectors $(\lambda_1, \ldots, \lambda_M)$ such that

$$0 \leq \lambda_j \leq \mu C_j \qquad 1 \leq j \leq M$$

and is therefore a compact set.

We wish to find a point which maximizes power over the feasible region. Since $\rho_j = \lambda_j / \mu C_j$ (for all $1 \leq j \leq M$), this optimization problem is equivalent to one involving the M unknowns $\rho_1, \ldots, \rho_M$. The feasible region of this equivalent problem is simply the (unit) M-cube given by the equations

$$0 \leq \rho_j \leq 1 \qquad 1 \leq j \leq M$$

(the cases $M = 2$ and $M = 3$ are depicted in Figure 3.2).

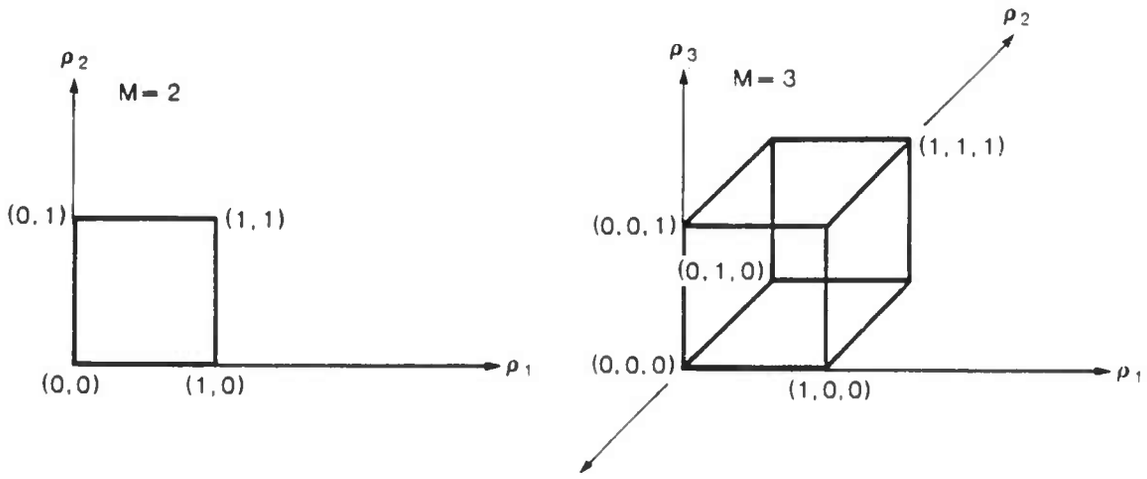

Figure 3.2 The Feasible Region

The power function P written in terms of the ρ_i is

$$P = \frac{\left(\sum_{i=1}^{M} \mu C_i \rho_i\right)^2}{\sum_{i=1}^{M} \frac{\rho_i}{1-\rho_i}}$$

We will frequently skip back and forth between these two equivalent formulations of this maximization problem. The feasible region is compact, and the objective function (the power function) is continuous there. Thus the function does attain its maximum, and it makes sense to try to find a globally optimal point.

3.2 A Bit of Optimization Theory

Let us first review some well-known facts in the classical theory of the optimization of real-valued functions of more than one variable. These results are standard and may be found, for example, in the books by Luenberger [Luen73] and Marlow [Marl78], among others. We consider a function $f: S \rightarrow R$ where S is an open set and $S \subseteq R^n$. Analogous to the derivative of a real-valued function of a single variable used in chapter 2, we first recall the concept of the *gradient* of the function f (denoted ∇f). This is simply the vector of partial derivatives

$$\nabla f = \left(\frac{\partial f}{\partial x_1}, \ldots, \frac{\partial f}{\partial x_n}\right)$$

We next recall the definition of a feasible direction. Consider a set $F \subseteq S$ (we think of F as being the feasible points for some set of constraints). A non-zero vector $\mathbf{d} \in R^n$ is called a *feasible direction* at $\mathbf{x} \in F$ if there exists $\beta > 0$ such that $\mathbf{x} + \alpha\mathbf{d} \in F$ whenever $\alpha \in [0, \beta]$. Also recall that if $\mathbf{d}$ is a unit vector in R^n, then the *directional derivative* of f in the direction $\mathbf{d}$ may be shown to be the following dot product

$$\nabla f \cdot \mathbf{d} = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i} d_i$$

Using these concepts, we now state a well-known result of optimization theory.

Fact: If a point $\mathbf{x} \in F$ is a local maximum (local minimum) of the function f on the set F, then $\nabla f \cdot \mathbf{d} \leq 0$ ($\nabla f \cdot \mathbf{d} \geq 0$) for any feasible direction $\mathbf{d}$ at the point $\mathbf{x}$.

The above result easily yields that $\nabla f = 0$ at a local maximum (or local minimum) of f which is an interior point of F, because all directions at such a point are feasible directions. We adopt a common definition in mathematical optimization and call a point $\mathbf{x}$ satisfying $\nabla f = 0$ a *critical point* of the function f.

We may apply these results to our optimization problem by observing that, in order to find the global maximizer of the power function P over the feasible region, we must determine all critical points of P (points with $\nabla P = 0$) and compare them to the optimal boundary point. (Recall that in the single variable case the boundary points yielded zero power, so that the optimal solution occurred where the derivative was zero.)

3.3 Characterization of the Optimal Solution

Let us now examine the characteristics of the optimal boundary point of our optimization problem. We adopt the convention of listing the channels of a parallel network in order of decreasing capacity. That is, we renumber the channels of our parallel configuration so that $C_1 \geq C_2 \geq \cdots \geq C_M$. This will not affect our optimization problem and will simplify subsequent notation. We now observe the following (intuitively clear) result which is easily shown.

Lemma: Consider two parallel M/M/1 systems, each with M channels, such that $C_{1i} \geq C_{2i}$ for $1 \leq i \leq M$ (that is, each channel of the first system has higher capacity than the corresponding channel of the second system). Then the optimal value of power for the first system is at least as large as the optimal value of power for the second system, i.e., $P_1^* \geq P_2^*$.

To prove this, let $(\lambda_1, \ldots, \lambda_M)$ be any point which is feasible for the second system. Then it is also feasible for the first system, and we have

$$P_1(\lambda_1, \ldots, \lambda_M) = \frac{\left(\sum_{i=1}^{M} \lambda_i\right)^2}{\sum_{i=1}^{M} \lambda_i T_{1i}} \geq \frac{\left(\sum_{i=1}^{M} \lambda_i\right)^2}{\sum_{i=1}^{M} \lambda_i T_{2i}} = P_2(\lambda_1, \ldots, \lambda_M)$$

because (for M/M/1)

$$T_{1i} = \frac{1}{\mu C_{1i} - \lambda_i} \leq \frac{1}{\mu C_{2i} - \lambda_i} = T_{2i}$$

for $1 \leq i \leq M$. Therefore, if we let $(\lambda_{21}^*, \ldots, \lambda_{2M}^*)$ be optimal for the second system, it is feasible for the first system, and we have

$$P_2^* = P_2(\lambda_{21}^*, \ldots, \lambda_{2M}^*) \leq P_1(\lambda_{21}^*, \ldots, \lambda_{2M}^*) \leq P_1^*$$

which gives the result.

This simple fact allows us to determine the nature of the optimal boundary point. The boundary of the feasible region consists of $2M$ faces, each face corresponding to $\lambda_i = 0$ or $\lambda_i = \mu C_i$ $(1 \leq i \leq M)$. That is, the boundary of the feasible region F may be expressed as

$$\bigcup_{i=1}^{M} \{(\lambda_1, \ldots, \lambda_M) \in F : \lambda_i = 0\} \cup \bigcup_{i=1}^{M} \{(\lambda_1, \ldots, \lambda_M) \in F : \lambda_i = \mu C_i\}$$

We immediately disregard the faces $\lambda_i = \mu C_i$ (the power function is identically zero there since

the delay is infinite), leaving the M faces $\lambda_i = 0$ to consider. We next note that the value of the M channel power function on the face $\lambda_i = 0$ corresponds to that of a parallel system with $M-1$ channels where channel i (with capacity C_i) is dropped. Assuming, as always, that $C_1 \geq \cdots \geq C_M$, we apply the above lemma to observe that the optimal boundary point occurs on the face $\lambda_M = 0$. Thus, the optimal boundary point is obtained by considering a parallel system of $M-1$ channels with capacities $C_1 \geq \cdots \geq C_{M-1}$ (the channel with the lowest capacity being dropped) and optimizing this system with respect to its corresponding power function. Of course, the $M-1$ channel system may have its optimal solution at a boundary point of its feasible region. To find such a point, channel $M-1$ must be dropped, and a parallel system of $M-2$ channels with capacities $C_1 \geq \cdots \geq C_{M-2}$ must be optimized. This process continues until the optimal solution for the M channel system is found. Note that a different power function must be examined in each case. That is, when we consider the subnetwork consisting of only the first m channels with capacities $C_1 \geq \cdots \geq C_m$, the corresponding power function P^m for this subproblem is

$$P^m = \frac{\left(\sum_{i=1}^{m} \lambda_i\right)^2}{\sum_{i=1}^{m} \lambda_i T_i}$$

Thus we see that the optimal solution of the M channel parallel system, which we denote by $(\lambda_1^*, \ldots, \lambda_M^*)$, satisfies

$$\lambda_i^* > 0 \qquad 1 \leq i \leq M^*$$
$$\lambda_i^* = 0 \qquad M^* < i$$

where the index M^* satisfies $1 \leq M^* \leq M$ (if $M^* = M$ then all components of the optimizer are positive and the solution is an interior point). The λ_i^* for $1 \leq i \leq M^*$ are found by considering a parallel system with M^* channels and capacities $C_1 \geq \cdots \geq C_{M^*}$ and determining the critical points for the power function (P^{M^*}) of this system. Therefore, solving the optimization problem for the M/M/1 parallel network is reduced to finding critical points of power functions of the original network and certain subnetworks.

3.4 Determination of Critical Points

We now find points of an M/M/1 parallel system with M channels such that $\nabla P = 0$. Taking the partial derivative of P with respect to λ_j (where $1 \leq j \leq M$) in equation (3.1) yields

$$\frac{\partial P}{\partial \lambda_j} = \frac{2\left(\sum_{i=1}^{M} \lambda_i T_i\right)\left(\sum_{i=1}^{M} \lambda_i\right) - \left(\sum_{i=1}^{M} \lambda_i\right)^2 \left(\lambda_j \frac{dT_j}{d\lambda_j} + T_j\right)}{\left(\sum_{i=1}^{M} \lambda_i T_i\right)^2} \tag{3.2}$$

or

$$\frac{\partial P}{\partial \lambda_j} = \left[\frac{\sum_{i=1}^{M} \lambda_i}{\sum_{i=1}^{M} \lambda_i T_i}\right]^2 \left[\frac{2\sum_{i=1}^{M} \lambda_i T_i}{\sum_{i=1}^{M} \lambda_i} - \left(\lambda_j \frac{dT_j}{d\lambda_j} + T_j\right)\right] \tag{3.3}$$

If we assume that $\nabla P = 0$, then $\frac{\partial P}{\partial \lambda_j} = 0$ for all $j = 1, \ldots, M$, and thus we have

$$\lambda_j \frac{dT_j}{d\lambda_j} + T_j = \frac{2\sum_{i=1}^{M} \lambda_i T_i}{\sum_{i=1}^{M} \lambda_i} = \lambda_k \frac{dT_k}{d\lambda_k} + T_k \qquad \forall j, k \tag{3.4}$$

We note that this equation holds for a G/G/1 parallel net. However, in the general G/G/1 case, closed-form expressions for the delay functions T_j (let alone expressions for the derivatives $dT_j/d\lambda_j$) are unknown. Thus we restrict our attention to more analytically tractable queueing systems. Since we will ultimately analyze networks of M/M/1 queues as models of general computer networks, a particular choice of interest is the M/M/1 case. This turns out to be a felicitous choice, since for M/M/1 the equations (3.4) become linear and quadratic (and thus more easily solved) as we shall see later in this section. For these reasons, in the following we specialize to the case of an M/M/1 parallel network.

We now wish to rewrite equations (3.4) in terms of ρ_j, the equivalent variables of optimization (it turns out to be easier to work with ρ_j than with λ_j). We first express the mean delay T_j and the derivatives $dT_j/d\lambda_j$ as functions of λ_j. Since we assume each channel acts as an M/M/1 queueing system, we have (for $1 \leq j \leq M$)

$$T_j = \frac{1}{\mu C_j - \lambda_j}$$

and so

$$\frac{dT_j}{d\lambda_j} = \frac{1}{(\mu C_j - \lambda_j)^2} = T_j^2$$

Therefore,

$$\lambda_j \frac{dT_j}{d\lambda_j} + T_j = \frac{\lambda_j}{(\mu C_j - \lambda_j)^2} + \frac{1}{\mu C_j - \lambda_j}$$

or

$$\lambda_j \frac{dT_j}{d\lambda_j} + T_j = \frac{\mu C_j}{(\mu C_j - \lambda_j)^2}$$

Since the efficiency of the jth channel is $\rho_j = \lambda_j / \mu C_j$, substituting this relationship into the previous equation gives

$$\lambda_j \frac{dT_j}{d\lambda_j} + T_j = \frac{1}{\mu C_j} \cdot \frac{1}{(1-\rho_j)^2} \tag{3.5}$$

As discussed above, we may consider the variables of our optimization problem to be $\rho_1, \ldots, \rho_M$ (instead of $\lambda_1, \ldots, \lambda_M$). We have that $\nabla P = 0$ implies

$$\frac{1}{\mu C_j} \cdot \frac{1}{(1-\rho_j)^2} = \frac{1}{\mu C_k} \cdot \frac{1}{(1-\rho_k)^2} \qquad \forall j, k \tag{3.6}$$

or

$$(1-\rho_j)^2 C_j = (1-\rho_k)^2 C_k \qquad \forall j, k \tag{3.7}$$

Taking square roots yields

$$(1-\rho_j)\sqrt{C_j/C_k} = 1-\rho_k \qquad \forall j, k \tag{3.8}$$

This gives us $M-1$ independent *linear* equations in the M unknowns $\rho_1, \ldots, \rho_M$.

To determine an Mth independent equation, recall from equation (3.4) that

$$\lambda_j \frac{dT_j}{d\lambda_j} + T_j = \frac{2\sum_{i=1}^{M} \lambda_i T_i}{\sum_{i=1}^{M} \lambda_i} \qquad 1 \le j \le M$$

and so, using equation (3.5) and Little's result, we have

$$\frac{1}{\mu C_j} \cdot \frac{1}{(1-\rho_j)^2} = \frac{2\sum_{i=1}^{M} \bar{N}_i}{\sum_{i=1}^{M} \lambda_i} \qquad 1 \le j \le M$$

Multiplying the jth equation by λ_j and recalling that $\rho_j = \lambda_j/\mu C_j$ yields

$$\frac{\rho_j}{(1-\rho_j)^2} = \lambda_j \cdot \frac{2\sum_{i=1}^{M} \bar{N}_i}{\sum_{i=1}^{M} \lambda_i} \qquad 1 \le j \le M$$

Summing these M equations gives

$$\sum_{i=1}^{M} \frac{\rho_i}{(1-\rho_i)^2} = 2\sum_{i=1}^{M} \bar{N}_i \tag{3.9}$$

Since we are considering the ith channel as an M/M/1 system, then $\bar{N}_i = \rho_i/(1-\rho_i)$ and so

$$\sum_{i=1}^{M} \frac{\rho_i}{(1-\rho_i)^2} = 2\sum_{i=1}^{M} \frac{\rho_i}{1-\rho_i}$$

We therefore have

$$\sum_{i=1}^{M} \frac{\rho_i - 2\rho_i + 2\rho_i^2}{(1-\rho_i)^2} = 0$$

or

$$\sum_{i=1}^{M} \frac{2\rho_i^2 - \rho_i}{(1-\rho_i)^2} = 0$$

Multiplying the numerator and denominator of the ith term by C_i yields

$$\sum_{i=1}^{M} \frac{(2\rho_i^2 - \rho_i)C_i}{(1-\rho_i)^2 C_i} = 0$$

By equation (3.7) we see that the denominators of all M terms in the above sum are identical. Multiplying both sides of the above equality by this common value finally gives our Mth independent equation

$$\sum_{i=1}^{M} (2\rho_i^2 - \rho_i)C_i = 0 \tag{3.10}$$

This equation may also be written as

$$\sum_{i=1}^{M} \rho_i\left(\rho_i - \frac{1}{2}\right)C_i = 0 \tag{3.11}$$

Therefore, to find points $(\rho_1, \ldots, \rho_M)$ such that $\nabla P = 0$, we must solve M equations in M unknowns. We note that $M-1$ of the equations are linear (equation (3.8)), while the Mth equation is quadratic (equation (3.10)). Our solution strategy is to use the $M-1$ linear equations to eliminate $\rho_2, \ldots, \rho_M$ (i.e., express each in terms of ρ_1) and then use the Mth quadratic equation to solve for ρ_1. Thus we could have 0, 1, or 2 points which are critical points of P, depending on the given parameters (the channel capacities C_i).

3.5 Characteristics of Critical Points

Let us examine these M equations in greater detail to obtain some interesting characteristics of critical points of P (and thus of the optimal power point, since it is a critical point of the power function of a parallel subsystem with M^* channels). From equation (3.8) and the convention that $C_j \geq C_k$ for $1 \leq j \leq k \leq M$, we see that $1-\rho_j \leq 1-\rho_k$ for all $j \leq k$ and so

$$\rho_1 \geq \rho_2 \geq \cdots \geq \rho_M \tag{3.12}$$

for any point of the M channel parallel system where the gradient of the power function is zero. Recalling the nature of the optimal power point of this optimization problem, we see that $(\rho_1^*, \ldots, \rho_M^*)$ (the optimal solution of the M channel parallel system) satisfies

$$\rho_1^* \geq \rho_2^* \geq \cdots \geq \rho_M^* \tag{3.13}$$

and also

$$\rho_i^* > 0 \quad 1 \leq i \leq M^*$$
$$\rho_i^* = 0 \quad M^* < i$$

where the ρ_i^* are determined by deleting all channels with index $i > M^*$ and solving for $\nabla P = 0$ of this reduced network. Equation (3.13) agrees with our intuition that, in order to decrease delay T (and thus increase power P), faster channels should be more heavily utilized than slower channels.

We may also obtain a bound on the value of the utilization of the fastest channel at maximum power by first examining the value of ρ_1 for a critical point of the power function P. We note that, if $\rho_1 < 1/2$, then by equation (3.12), $\rho_i < 1/2$ for all $1 \leq i \leq M$. This is impossible by equation (3.11), and we thus conclude that

$$\rho_1 \geq \frac{1}{2} \tag{3.14}$$

for such a critical point. Therefore, since the optimal power point is a critical point of the power function P^{M^*}, we have

$$\rho_1^* \geq \frac{1}{2} \tag{3.15}$$

and so

$$\bar{N}_1^* \geq 1 \tag{3.16}$$

We now derive an interesting expression which has appeared before in several of our single-variable optimization problems and which will appear later in a more general network setting. Using equation (3.9) and the fact that $\rho_i/(1-\rho_i) = \bar{N}_i$ for an M/M/1 system, we have

$$\sum_{i=1}^{M} \frac{\bar{N}_i}{1-\rho_i} = 2\sum_{i=1}^{M} \bar{N}_i$$

or

$$\sum_{i=1}^{M} \bar{N}_i \left[2 - \frac{1}{1-\rho_i}\right] = 0$$

This becomes,

$$\sum_{i=1}^{M} \bar{N}_i \left[\frac{1-2\rho_i}{1-\rho_i}\right] = 0$$

which is equivalent to

$$\sum_{i=1}^{M} \bar{N}_i \left[1 - \frac{\rho_i}{1-\rho_i} \right] = 0$$

Again, we recall that $\rho_i/(1-\rho_i) = \bar{N}_i$, and so

$$\sum_{i=1}^{M} \bar{N}_i (1 - \bar{N}_i) = 0$$

or

$$\sum_{i=1}^{M} \bar{N}_i = \sum_{i=1}^{M} (\bar{N}_i)^2$$

The above equation holds for critical points of the power function of an M channel M/M/1 parallel network. We may once again exploit the nature of the global optimizer of the power function (it is a critical point for an M^* channel net) to write

$$\sum_{i=1}^{M^*} \bar{N}_i^* = \sum_{i=1}^{M^*} (\bar{N}_i^*)^2 \tag{3.17}$$

In fact, since $\rho_i^* = 0$ (and thus $\bar{N}_i^* = 0$) for $i > M^*$, we may actually regard the summations in equation (3.17) as including terms from 1 to M, which then yields the following

Theorem 3.1

For the M/M/1 parallel network (with unknown routing), the average number in system at maximum power satisfies

$$\bar{N}^* = \sum_{i=1}^{M} \bar{N}_i^* = \sum_{i=1}^{M} (\bar{N}_i^*)^2 \tag{3.18}$$

We note that this is identical to an equation found for the M/M/1 series net and also for the parallel M/M/1 net (where the routing was assumed to be known). As discussed above, this equation is a special case of a general result to be derived later in this work.

In the same manner as was done previously for the M/M/1 series net and the M/M/1 parallel net (with known routing) in chapter 2, we may use equation (3.18) to yield

Theorem 3.2

For the M/M/1 parallel network (with unknown routing), the average number in system at maximum power satisfies

$$\sum_{i=1}^{M} (1 - \bar{N}_i^*) = \sum_{i=1}^{M} (1 - \bar{N}_i^*)^2 \tag{3.19}$$

This interesting dual equation will be shown to hold in a more general setting in chapter 4.

As in chapter 2, equation (3.18) also yields the following

Theorem 3.3

For the M/M/1 parallel network (with unknown routing), the average number in system at maximum power satisfies

$$\overline{N}^* \leq M \tag{3.20}$$

This bound will be shown to hold for more general networks in chapter 4. Of course, we actually have the tighter bound

$$\overline{N}^* \leq M^* \tag{3.21}$$

from equation (3.17) (any channel j assigned an input rate $\lambda_j^* = 0$ does not contribute to $\overline{N}^*$). The index M^* is a quantity which is determined from the optimal power point (and thus is not very easy to calculate), while M is simply the number of channels in the network. Unfortunately, in some cases, $M^* << M$, and so equation (3.20) may not give a very good bound. Later in this chapter (in section 3.7.1) we describe one such case.

3.6 Solution of the Power Problem

Before solving the M independent equations, let us introduce some simplifying notation. To this end, we first define, for $1 \leq j \leq M$,

$$S_j \triangleq \sqrt{C_j/C_1} \tag{3.22}$$

Note that $1 = S_1 \geq S_2 \geq \cdots \geq S_M$. We also define, for $1 \leq j \leq M$,

$$\theta_j \triangleq 1 - \rho_j \tag{3.23}$$

Choosing $k = 1$ in equation (3.8), we have (for $1 \leq j \leq M$)

$$1 - \rho_1 = (1 - \rho_j)\sqrt{C_j/C_1}$$

which is, using our new terminology,

$$\theta_j = \frac{\theta_1}{S_j} \tag{3.24}$$

We also may utilize this notation in observing an interesting relationship among mean system times T_j at the M different channels for critical points of P. From equation (3.6) and the fact that $T_j = 1/(\mu C_j - \lambda_j)$ for M/M/1, we have

$$\frac{T_j}{1 - \rho_j} = \frac{T_k}{1 - \rho_k} \qquad \forall\, j, k$$

Therefore, choosing $k = 1$ in the above expression, we find

$$T_j = T_1\left(\frac{\theta_j}{\theta_1}\right) \qquad 1 \le j \le M$$

Using equation (3.24), we finally have

$$T_j = \frac{T_1}{S_j} \qquad 1 \le j \le M \tag{3.25}$$

which determines the average times at the channels in terms of the given capacities for critical points of P.

We now solve the M independent equations (in equations (3.8) and (3.10)) for $\rho_1, \ldots, \rho_M$. Equation (3.10) yields

$$\sum_{i=1}^{M}(2\rho_i - 1)\rho_i C_i = 0$$

which becomes, using the definition of θ_i,

$$\sum_{i=1}^{M}[2(1-\theta_i) - 1](1-\theta_i)C_i = 0$$

or

$$\sum_{i=1}^{M}(1-2\theta_i)(1-\theta_i)C_i = 0$$

Dividing by C_1 and recalling that $C_i/C_1 = S_i^2$, we have

$$\sum_{i=1}^{M}(1-2\theta_i)(1-\theta_i)S_i^2 = 0$$

or

$$\sum_{i=1}^{M}(1-3\theta_i+2\theta_i^2)S_i^2 = 0$$

Using equation (3.24), we obtain

$$\sum_{i=1}^{M}\left[1-3\frac{\theta_1}{S_i}+2\left(\frac{\theta_1}{S_i}\right)^2\right]S_i^2 = 0$$

or

$$\sum_{i=1}^{M}(2\theta_1^2 - 3S_i\theta_1 + S_i^2) = 0$$

Having eliminated $\theta_2, \ldots, \theta_M$, we finally have the following quadratic equation in θ_1:

$$2M\theta_1^2 - 3\left(\sum_{i=1}^{M}S_i\right)\theta_1 + \sum_{i=1}^{M}S_i^2 = 0 \tag{3.26}$$

The two roots of this equation are

$$u \triangleq \frac{3\sum_{i=1}^{M} S_i + \left[9\left(\sum_{i=1}^{M} S_i\right)^2 - 8M\sum_{i=1}^{M} S_i^2\right]^{\frac{1}{2}}}{4M} \tag{3.27}$$

and

$$v \triangleq \frac{3\sum_{i=1}^{M} S_i - \left[9\left(\sum_{i=1}^{M} S_i\right)^2 - 8M\sum_{i=1}^{M} S_i^2\right]^{\frac{1}{2}}}{4M} \tag{3.28}$$

These roots are real if and only if we have a non-negative discriminant, i.e.,

$$D \triangleq 9\left(\sum_{i=1}^{M} S_i\right)^2 - 8M\sum_{i=1}^{M} S_i^2 \geq 0 \tag{3.29}$$

In this case, we clearly have $0 < v \leq u$.

We now claim that

$$u \leq 2v \tag{3.30}$$

Using equations (3.27) and (3.28), we see that equation (3.30) holds if and only if

$$3\sum_{i=1}^{M} S_i + \sqrt{D} \leq 6\sum_{i=1}^{M} S_i - 2\sqrt{D}$$

which is equivalent to

$$\sqrt{D} \leq \sum_{i=1}^{M} S_i$$

Squaring this last equation and using the definition of D given in equation (3.29), we obtain

$$\left(\sum_{i=1}^{M} S_i\right)^2 \leq M\sum_{i=1}^{M} S_i^2 \tag{3.31}$$

We note that equation (3.31) is true by using the Cauchy-Schwarz inequality

$$\left(\sum_i x_i y_i\right)^2 \leq \left(\sum_i x_i^2\right)\left(\sum_i y_i^2\right)$$

with

$$x_i = 1$$

and

$$y_i = S_i$$

This shows equation (3.30). Therefore, whenever the roots u and v are real (i.e., whenever

equation (3.29) holds), we have the bounds

$$0 < v \leq u \leq 2v \tag{3.32}$$

Assuming the root u is real, let us see when it leads to a feasible solution. We denote the point corresponding to u as $\rho(u)$ (or $\lambda(u)$ or $\theta(u)$, depending on the variable of interest), but suppress the dependence on the particular root for notational convenience when it is clear from the context. We call the root u feasible if the corresponding point $\rho(u)$ is feasible. The condition for feasibility is $0 \leq \rho_j < 1$ for $1 \leq j \leq M$. For the point $\theta(u)$, we have $\theta_1 = u$, and from equation (3.24), $\theta_j = u/S_j$ for $1 \leq j \leq M$. Equation (3.32) then gives $0 < \theta_j$ for all j, which is equivalent to $\rho_j < 1$. Thus the solution is feasible if and only if $\rho_j \geq 0$ for $1 \leq j \leq M$. This is equivalent to $\theta_j \leq 1$ for all j, or $u \leq S_j$. Since $S_j \geq S_k$ for $j \leq k$, the condition for feasibility becomes

$$u \leq S_M$$

(note that this is equivalent to $\rho_M \geq 0$). The feasible point is an interior point if and only if $u < S_M$ ($\rho_M > 0$). Similarly, the root v would give a feasible point if and only if

$$v \leq S_M$$

with the feasible point an interior point if and only if $v < S_M$. If these roots (u and v) are real and give rise to solutions within the feasible region, then they give values of θ_1 where the gradient of the power function is zero. Using the relationship $\theta_j = \theta_1/S_j$ where $\theta_j = 1 - \rho_j$, all points $(\rho_1, \ldots, \rho_M)$ which satisfy $\nabla P = 0$ for the M channel parallel network may be found. As described above, solutions of this form (for the original network and certain subnetworks) will lead to the global maximizer of P over the region of feasibility.

Recall our assumption that both u and v are real (and thus equation (3.32) holds). If $S_M < v$, then neither root yields a feasible solution. If $v \leq S_M < u$, then only v yields a feasible solution, and we may consequently disregard u. We now show that if $u \leq S_M$ so that both u and v yield feasible solutions, then the solution corresponding to v gives higher power P. Therefore, the root u need not be considered in determining optimal power points. To this end, assume u and v are real and the solution points corresponding to them are both feasible. We examine the solution given by the root v. For $1 \leq i \leq M$ we have

$$\overline{N}_i = \frac{\rho_i}{1-\rho_i} = \frac{1-\theta_i}{\theta_i} = \frac{1-v/S_i}{v/S_i}$$

or

$$\overline{N}_i = \frac{S_i - v}{v}$$

We also have

$$\lambda_i = \mu C_i \rho_i = \mu C_1 \cdot \frac{C_i}{C_1}(1 - \theta_i) = \mu C_1 \cdot S_i^2 (1 - v/S_i)$$

or

$$\lambda_i = \mu C_1 S_i (S_i - v)$$

Therefore, the value of power at this feasible point (call the value $P(v)$, instead of $P(\rho(v))$, by abuse of notation) is

$$P(v) = \frac{\left(\sum_{i=1}^{M} \lambda_i\right)^2}{\sum_{i=1}^{M} \overline{N}_i} = \frac{(\mu C_1)^2 \left[\sum_{i=1}^{M} S_i (S_i - v)\right]^2}{\frac{1}{v} \sum_{i=1}^{M} (S_i - v)}$$

or

$$P(v) = (\mu C_1)^2 v \frac{\left[\sum_{i=1}^{M} S_i^2 - v \sum_{i=1}^{M} S_i\right]^2}{\sum_{i=1}^{M} S_i - Mv}$$

We now use the well-known result that if y and z are the two roots of the quadratic equation $ax^2 + bx + c = 0$, then $y + z = -b/a$ and $y \cdot z = c/a$. Applying this to the roots u and v of equation (3.26), we find that

$$u + v = \frac{3 \sum_{i=1}^{M} S_i}{2M}$$

and

$$u \cdot v = \frac{\sum_{i=1}^{M} S_i^2}{2M}$$

We therefore have

$$\sum_{i=1}^{M} S_i = \frac{2M}{3}(u + v)$$

and

$$\sum_{i=1}^{M} S_i^2 = 2M(u \cdot v)$$

Using these expressions in the equation for $P(v)$ gives

$$P(v) = (\mu C_1)^2 v \frac{\left[2Muv - v\frac{2M}{3}(u+v)\right]^2}{\frac{2M}{3}(u+v) - Mv}$$

or

$$P(v) = (\mu C_1)^2 v \frac{4M^2v^2\left(\frac{2}{3}u - \frac{1}{3}v\right)^2}{M\left(\frac{2}{3}u - \frac{1}{3}v\right)}$$

We finally have

$$P(v) = (\mu C_1)^2 \frac{4M}{3} v^3(2u - v) \tag{3.33}$$

In a similar fashion, we examine the solution corresponding to u (call its value $P(u)$) and find that

$$P(u) = (\mu C_1)^2 \frac{4M}{3} u^3(2v - u) \tag{3.34}$$

We claim that $P(v) - P(u) \geq 0$. This difference is nonnegative if and only if

$$v^3(2u - v) - u^3(2v - u) \geq 0$$

which we show as follows. We have

$$\begin{aligned} v^3(2u - v) - u^3(2v - u) &= u^4 - v^4 - 2u^3v + 2uv^3 \\ &= (u^2 + v^2)(u^2 - v^2) - 2uv(u^2 - v^2) \\ &= (u^2 - v^2)(u - v)^2 \\ &= (u + v)(u - v)^3 \end{aligned}$$

which is nonnegative because $0 < v \leq u$ whenever the roots are real. Therefore $P(v) \geq P(u)$, and we consequently may disregard the root u in determining the optimal power point.

Let us now review the solution procedure for finding the global maximizer of power for an M/M/1 parallel network with M channels and channel capacities $C_1 \geq \cdots \geq C_M$. Let P^m $(1 \leq m \leq M)$ be the power function for the subnetwork consisting of the fastest m channels (channel capacities $C_1, \ldots, C_m$), and let u_m and v_m be the corresponding roots when P^m is used. We first calculate $v = v_M$ from equation (3.28). If v_M is real and $v_M \leq S_M$, then v_M corresponds to a feasible point $\rho(v_M) = (\rho_1(v_M), \ldots, \rho_M(v_M))$, which is a critical point of the original power function $P = P^M$. The coordinates of this critical point are given by $\rho_i(v_M) = 1 - v_M/S_i$ for $1 \leq i \leq M$, and the value of P^M at this point is given by equation

(3.33) where $u = u_M$ satisfies equation (3.27). This procedure is then carried out for subnetworks obtained from the original network by dropping the slowest channel in each case. For example, v_{M-1} (which corresponds to a critical point $\rho(v_{M-1})$ of P^{M-1}) and u_{M-1} are calculated next. Among these M points (corresponding to the roots v_j, $1 \leq j \leq M$, of the M subnetworks), that one which is feasible and has the highest power is the global maximizer for our optimization problem. The following illustrates the procedure.

Step 0: Set $m = M$

Step 1: Calculate v_m (and u_m)

Step 2: Check if v_m real; if not, go to Step 5

Step 3: Check if v_m feasible; if not, go to Step 5

Step 4: Evaluate $P^m(v_m) = P^M(v_m)$

Step 5: Check if $m = 1$; if not, set $m = m - 1$ and go to Step 1

Step 6: Determine the global optimum from the solutions calculated in Step 4

Thus, in general, we must search over M points, although later in this chapter (in section 3.7) we obtain conditions on the input parameters (the capacities C_i) which reduce the number of points that we need to examine.

3.6.1 Equal Capacity Case

We now consider several examples in greater detail. For our first (simple) example, let us assume that all channels have the same capacity C. In this case, we have $S_j = 1$ for all $1 \leq j \leq M$, so that the root $v = v_M$ becomes

$$v = \frac{3M - \sqrt{9M^2 - 8M^2}}{4M} = \frac{3M - M}{4M} = \frac{1}{2}$$

From equation (3.24) we must have $\theta_j = \theta_k$ for all j, k, so that the root v_M yields

$$\theta_j = \frac{1}{2} \qquad 1 \leq j \leq M$$

We also have $\rho_j = \rho_k$, $\bar{N}_j = \bar{N}_k$, and $\lambda_j = \lambda_k$ for all j, k. Thus

$$\rho_j = \frac{1}{2}, \quad \bar{N}_j = 1, \quad \lambda_j = \frac{\mu C}{2} \qquad 1 \leq j \leq M$$

and the solution given by v_M is $\rho(v_M) = (1/2, \ldots, 1/2)$. Therefore (writing $P = P^M$)

$$P(v_M) = P^M(v_M) = \frac{(M \mu C/2)^2}{M} = M\left(\frac{\mu C}{2}\right)^2$$

We see that this must be the global maximum as follows. For any $1 \leq m < M$, the root v_m clearly yields the solution $\rho(v_m) = (1/2, \ldots, 1/2, 0, \ldots, 0)$ where the first m components are $1/2$ and the last $M - m$ components are zero. This has power

$$P(v_m) = P^M(v_m) = P^m(v_m) = m\left(\frac{\mu C}{2}\right)^2$$

and so the solution corresponding to v_M is globally optimal. Thus we have $\overline{N}_j^* = 1$ for all j, and so this equal capacity parallel channel case generalizes the "keep the pipe full" result of Kleinrock [Klei78a] for the M/M/1 single-node system.

3.6.2 The Two Channel Case

The next case we shall consider is that of a network having only two parallel channels with capacities $C_1 \geq C_2$ (that is, the case $M = 2$). This simple net was examined by Jaffe [Jaff81], whose contributions we will discuss at the end of this subsection. We obtain a complete analytical solution to this system by finding the input rates which optimize power for all possible values of the given channel capacities C_1 and C_2.

Following the procedure discussed at the end of section 3.6 (with $M = 2$), we first find the solution corresponding to v_2 (the case $m = 2$) and then find the solution corresponding to v_1 (the case $m = 1$). Note that v_2 (if it is real and feasible) yields the optimal critical point of $P = P^2$, while v_1 (which is always real and feasible) yields the optimal boundary point. For notational convenience, define $S \triangleq S_2 \left(= \sqrt{C_2/C_1}\right)$; thus we have $0 < S \leq 1$. We look for the optimal point satisfying $\nabla P = 0$ (which is given by $v = v_2$) by examining equation (3.28) for this network. The root v is

$$v = \frac{3(1+S) - \sqrt{D}}{8}$$

where

$$D = 9(1+S)^2 - 16(1+S^2) = 9 + 18S + 9S^2 - 16 - 16S^2$$

Thus we have

$$D = -7S^2 + 18S - 7$$

and so

$$v = \frac{3 + 3S - \sqrt{-7S^2 + 18S - 7}}{8}$$

In order for this root to be real, we must have $D = -7S^2 + 18S - 7 \geq 0$. This is equivalent to

$$S^2 - \frac{18}{7}S + 1 \leq 0$$

or

$$\left(S - \frac{9}{7}\right)^2 \leq \frac{32}{49}$$

Thus the condition for v to be real becomes

$$-\frac{4\sqrt{2}}{7} \leq S - \frac{9}{7} \leq \frac{4\sqrt{2}}{7}$$

or

$$\frac{9-4\sqrt{2}}{7} \leq S \leq \frac{9+4\sqrt{2}}{7}$$

Since $S \leq 1$, we must have

$$\frac{9-4\sqrt{2}}{7} \leq S \leq 1$$

These inequalities can be written as (approximately) $.47759 \leq S \leq 1$. If this condition is not satisfied, P has no critical point; hence the boundary point given by v_1 is globally optimal.

Under the above restrictions on the given parameter S, let us examine the root $v = v_2$ and determine if it leads to a feasible solution. We have $\rho_1 = 1 - v$, so that

$$\rho_1 = \frac{5 - 3S + \sqrt{-7S^2 + 18S - 7}}{8} \tag{3.35}$$

Now $\theta_2 = v/S$ and $\rho_2 = 1 - \theta_2$, so that

$$\rho_2 = \frac{5S - 3 + \sqrt{-7S^2 + 18S - 7}}{8S} \tag{3.36}$$

Recalling the discussion of feasibility after equation (3.28), we see that v yields a feasible point if and only if $\rho_2 \geq 0$. Clearly, ρ_2 is non-negative for $S \geq 3/5$, so we examine the case that $S < 3/5$. Then in order for $\rho_2 \geq 0$, we must have

$$\sqrt{-7S^2 + 18S - 7} \geq 3 - 5S$$

Since both sides of this inequality are nonnegative in the range of S under consideration, squaring will preserve it, and our condition becomes

$$-7S^2 + 18S - 7 \geq 9 - 30S + 25S^2$$

or

$$32S^2 - 48S + 16 \leq 0$$

Factoring yields

$$16(2S - 1)(S - 1) \leq 0$$

which is seen to hold (in the range under consideration) when $1/2 \leq S < 3/5$. Thus v yields a feasible point ($\rho_2 \geq 0$) for $1/2 \leq S \leq 1$. Putting together the previous cases, the root v is

real and yields a feasible point (ρ_1, ρ_2) which is a critical point of P if and only if S is in the range $1/2 \leq S \leq 1$. The values for ρ_1 and ρ_2 are given above in equations (3.35) and (3.36).

We next consider the case $m = 1$ of our optimization procedure. To find the optimal boundary point (given by v_1), we must examine the face $\lambda_2 = 0$ (the slowest channel is dropped). This face corresponds to a single M/M/1 system with capacity C_1, and so, from the above-mentioned result of Kleinrock, we find that the optimal boundary point is $(\rho_1, \rho_2) = (1/2, 0)$.

We now determine the optimal power point (step 6 of our procedure), which depends on the given parameter S. Intuitively, we expect that for some constant K, if $C_1/C_2 > K$ then $\lambda_2^* = 0$. That is, if one channel is much faster than the other, any traffic which uses the slow channel will increase the mean system delay T significantly and thus greatly decrease the power. It would therefore seem to be an optimal strategy to allow no traffic on the slow channel; in fact, we see below that our intuition is correct (and also that $K = 4$).

Since an optimal interior point must have $\nabla P = 0$ (and thus corresponds to $v = v_2$), we immediately see that the global maximum occurs at the optimal boundary point when $S < 1/2$ (i.e., $C_1 > 4C_2$). For $S = 1/2$ (i.e., $C_1 = C_2$), the optimal boundary point and the critical point given by v coincide, and so this point must be the global maximum. Thus we have $(\rho_1^*, \rho_2^*) = (1/2, 0)$ when $C_1 \geq 4C_2$.

For $1/2 < S \leq 1$ $(C_2 \leq C_1 < 4C_2)$, we must compare the value of power at the point given by the root v satisfying $\nabla P = 0$ (which is an interior point for this range of S) with that from the optimal boundary point. Rather than finding the power itself (a difficult computation) we will make use of the optimization theory fact stated at the end of section 3.2 above. In order to determine the global maximizer for the range $1/2 < S \leq 1$, we examine the gradient of the power function at the optimal boundary point. Recall from equation (3.3) that (for $j = 1, 2$)

$$\frac{\partial P}{\partial \lambda_j} = \left[\frac{\lambda_1 + \lambda_2}{\lambda_1 T_1 + \lambda_2 T_2}\right]^2 \left[\frac{2(\lambda_1 T_1 + \lambda_2 T_2)}{\lambda_1 + \lambda_2} - \left(\lambda_j \frac{dT_j}{d\lambda_j} + T_j\right)\right]$$

Using the fact that $dT/d\lambda = T^2$ for M/M/1 along with Little's result, the jth partial derivative simplifies to

$$\frac{\partial P}{\partial \lambda_j} = \left[\frac{\lambda_1 + \lambda_2}{\bar{N}_1 + \bar{N}_2}\right]^2 \left[\frac{2(\bar{N}_1 + \bar{N}_2)}{\lambda_1 + \lambda_2} - T_j(\bar{N}_j + 1)\right]$$

At $(\rho_1, \rho_2) = (1/2, 0)$ we have $(\lambda_1, \lambda_2) = (\mu C_1/2, 0)$. At this optimal boundary point we also have $\bar{N}_1 = 1$, $\bar{N}_2 = 0$, $T_1 = 1/\lambda_1 = 2/\mu C_1$, and $T_2 = 1/\mu C_2 = 1/S^2 \mu C_1$ (by the definition of S). Therefore,

$$\frac{\partial P}{\partial \lambda_1} = \left(\frac{\mu C_1}{2}\right)^2 \left(\frac{4}{\mu C_1} - \frac{2}{\mu C_1} \cdot 2\right) = 0$$

which is as expected. Also,

$$\frac{\partial P}{\partial \lambda_2} = \left(\frac{\mu C_1}{2}\right)^2 \left(\frac{4}{\mu C_1} - \frac{1}{S^2 \mu C_1} \cdot 1\right)$$

or

$$\frac{\partial P}{\partial \lambda_2} = \mu C_1 \left[1 - \frac{1}{4S^2}\right]$$

Thus, at the optimal boundary point $(\lambda_1, \lambda_2) = (\mu C_1/2, 0)$, we find that the gradient of P has value

$$\nabla P = (0, \mu C_1[1 - 1/4S^2])$$

By the well-known result from optimization theory quoted above, if a point is a local maximum of P, then $\nabla P \cdot \mathbf{d} \leq 0$ for every feasible direction $\mathbf{d}$. Clearly, for $1/2 < S \leq 1$ the point $(\mu C_1/2, 0)$ cannot be even a local maximum because there are (infinitely many) feasible directions $\mathbf{d}$ with $\nabla P \cdot \mathbf{d} > 0$. In fact, at that point any vector $\mathbf{d} = (d_1, d_2)$ with $d_2 > 0$ is a feasible direction such that

$$\nabla P \cdot \mathbf{d} = \mu C_1 \left[1 - \frac{1}{4S^2}\right] d_2 > 0$$

Thus for $1/2 < S \leq 1$ the maximal boundary point is not globally optimal (it is not even locally optimal), and so the critical point given by the root v must be optimal. Writing S in terms of the given channel capacities, the above cases enable us to prove the following

Theorem 3.4

The optimal solution which maximizes the power of the two channel M/M/1 parallel network is:

(a) *for* $C_2 \leq C_1 \leq 4C_2$ *then*

$$\rho_1^* = \frac{5 - 3S + \sqrt{-7S^2 + 18S - 7}}{8}, \qquad \rho_2^* = \frac{5S - 3 + \sqrt{-7S^2 + 18S - 7}}{8S} \tag{3.37}$$

(b) *for* $4C_2 \leq C_1$ *then*

$$\rho_1^* = \frac{1}{2}, \qquad \rho_2^* = 0 \tag{3.38}$$

We now examine the behavior of the optimal solution for case (a) of Theorem 3.4 in greater detail; in particular, we focus on the parameter ρ_1^*. From equation (3.15), we know that $\rho_1^* \geq 1/2$ for all values of S; by direct calculation we also note that $\rho_1^* = 1/2$ for $S = 1/2$ and $S = 1$. We find below that ρ_1^* is a concave function of S for $1/2 \leq S \leq 1$ (i.e., case (a) of the theorem) and achieves a maximum value in this interval. Since $\rho_1^* = 1 - v$ for $1/2 \leq S \leq 1$, we may examine the root $v = v_2$ as a function of S in order to determine the behavior of ρ_1^*. Recall that

$$v = \frac{3(1+S) - \sqrt{D}}{8}$$

where

$$D = -7S^2 + 18S - 7$$

for the range of S under consideration (i.e., $1/2 \leq S \leq 1$). Differentiating v with respect to S yields

$$\frac{dv}{dS} = \frac{1}{8}\left[3 - \frac{9 - 7S}{\sqrt{D}}\right]$$

and (after a bit of manipulation)

$$\frac{d^2v}{dS^2} = \frac{4}{D^{3/2}}$$

Since $D > 0$ for $1/2 \leq S \leq 1$, we have $d^2v/dS^2 > 0$, and thus v is a strictly convex function of S for this range of S. Therefore, $\rho_1^* = 1 - v$ is a strictly concave function of S. To find the maximum value of ρ_1^*, we set $d\rho_1^*/dS = 0$, or equivalently $dv/dS = 0$. This gives

$$9 - 7S = 3\sqrt{D}$$

Squaring this equation and using the definition of D yields (after some calculation)

$$7S^2 - 18S + 9 = 0$$

Thus the value of S which gives the maximum ρ_1^* must satisfy the above quadratic equation. Note that the corresponding value of D satisfies

$$D = -7S^2 + 18S - 7 = 2$$

The two roots of the quadratic equation are $S = (9 \pm 3\sqrt{2})/7$. Since $S \leq 1$, we must choose the negative square root which gives

$$S = \frac{9 - 3\sqrt{2}}{7} \cong .679623$$

The corresponding value of v is

$$v = \frac{6 - 2\sqrt{2}}{7} \cong .453082$$

Note that $v = (2/3)S$. We now find that the maximum value of ρ_1^* is

$$\rho_1^* = 1 - v = \frac{1+2\sqrt{2}}{7} \cong .546918$$

while we also have

$$\rho_2^* = 1 - \frac{v}{S} = \frac{1}{3}$$

for this value of S which optimizes ρ_1^*.

The following table lists several important system parameters for various values of the variable S $(= \sqrt{C_2/C_1})$:

S	ρ_1^*	ρ_2^*	$\bar{N}_1^*$	$\bar{N}_2^*$	$\bar{N}^*$	$P^*/(\mu C_1)^2$
1	1/2	1/2	1	1	2	1/2
⋮	⋮	⋮	⋮	⋮	⋮	⋮
.6796	.5469	1/3	1.2071	1/2	1.7071	.2878
⋮	⋮	⋮	⋮	⋮	⋮	⋮
7/11	6/11	2/7	6/5	2/5	8/5	.2732
⋮	⋮	⋮	⋮	⋮	⋮	⋮
4/7	15/28	3/16	15/13	3/13	18/13	.2574
⋮	⋮	⋮	⋮	⋮	⋮	⋮
$\leq 1/2$	1/2	0	1	0	1	1/4

Table 3.1

Several interesting consequences of this simple two channel parallel network may be drawn from the above results. Let us regard the network as a model of two users (source-destination pairs), each with its own channel for its packets. Then we see that operating at the (globally) optimal power point may be unfair to some users in the sense that they are restricted to having zero throughput. In fact, from the above characterization of the global maximum of this system, we see that user 2 will have zero throughput whenever $4C_2 \leq C_1$ (i.e., whenever the faster channel is at least four times as fast as the slower channel). Of course, such a system operating point is unfair to user 2. We also note that local power is not necessarily equal to global power, i.e., an operating point obtained by maximizing power using only local information (where each user is aware only of traffic characteristics along his own path) may not be the operating point obtained by globally maximizing power. We see that the optimal point using local power information is $(\rho_1, \rho_2) = (1/2, 1/2)$, which is globally optimal if and only if $S = 1$

(i.e., if and only if both channels are the same speed). Thus, an algorithm which attempts to optimize power using only local information will, in general, fail.

As stated above, this two channel parallel net was examined by Jaffe [Jaff81] with regard to power. By means of a counterexample, he first demonstrated that the optimal local power point need not be globally optimal. Using this two channel net, Jaffe was also the first to study fairness issues when maximizing power. Here we have rigorously identified the behavior of the optimal power point for all C_i. In so doing, we have extended Jaffe's result in which he found the optimal solution for the special case of $C_1/C_2 \to \infty$ ($S \to 0$ in our terminology). Using this limiting case, Jaffe was the first to point out that unfair operating points exist which (globally) maximize power.

3.7 Simplifying the Determination of the Optimal Solution

In the two channel example analyzed above, the global optimum was obtained by examining the gradient of the power function at the optimal boundary point. Using this example as a guide, we find that it is sometimes possible to analytically determine the optimal solution without evaluating all the M roots v_j for $1 \le j \le M$. That is, we may eliminate the tail of our optimization procedure by stopping with root v_m for some index m.

Before we consider the general case, let us study the special case where only one root needs to be evaluated. We wish to find conditions on the parameters S_i (and thus the channel capacities C_i) which insure that the points corresponding to the roots $v_1, v_2, \ldots, v_{M-1}$ are not optimal. In such a case, our optimization procedure involves one iteration, and since the conditions we seek are in terms of the given S_i only, the remaining $M-1$ roots need not be evaluated. We proceed by examining the gradient of the power function for the original M channel network at these $M-1$ points. In particular, consider an index $1 \le m < M$ (which corresponds to a subnetwork with $m < M$ channels) and assume that v_m yields a feasible point which is a critical point of its power function P^m. That is, v_m is real and satisfies $v_m \le S_m$. The point $(\rho_1, \ldots, \rho_m)$ (or $(\lambda_1, \ldots, \lambda_m)$) which corresponds to the root v_m is given by $\rho_j = \rho_j(v_m) = 1 - v_m/S_j$ for $1 \le j \le m$ and satisfies

$$\lambda_j \frac{dT_j}{d\lambda_j} + T_j = \frac{2\sum_{i=1}^{m} \lambda_i T_i}{\sum_{i=1}^{m} \lambda_i} \qquad 1 \le j \le m \tag{3.39}$$

We now evaluate the gradient of $P = P^M$ at the point $(\lambda_1, \ldots, \lambda_m, 0, \ldots, 0)$. Note that this point represents the solution corresponding to the root v_m in the original M-dimensional space. Evaluating equation (3.3) for the jth partial derivative of $P = P^M$ at the above-mentioned point, we find (since $\lambda_j = 0$ for $m < j \le M$)

$$\frac{\partial P}{\partial \lambda_j} = \left[\frac{\sum_{i=1}^{m}\lambda_i}{\sum_{i=1}^{m}\lambda_i T_i}\right]^2 \left[\frac{2\sum_{i=1}^{m}\lambda_i T_i}{\sum_{i=1}^{m}\lambda_i} - \left(\lambda_j \frac{dT_j}{d\lambda_j} + T_j\right)\right]$$

From equation (3.39) we have

$$\lambda_1 \frac{dT_1}{d\lambda_1} + T_1 = \frac{2\sum_{i=1}^{m}\lambda_i T_i}{\sum_{i=1}^{m}\lambda_i} = \lambda_j \frac{dT_j}{d\lambda_j} + T_j \qquad 1 \le j \le m$$

and so (for $1 \le j \le M$)

$$\frac{\partial P}{\partial \lambda_j} = \left[\frac{\sum_{i=1}^{m}\lambda_i}{\sum_{i=1}^{m}\lambda_i T_i}\right]^2 \left[\left(\lambda_1 \frac{dT_1}{d\lambda_1} + T_1\right) - \left(\lambda_j \frac{dT_j}{d\lambda_j} + T_j\right)\right]$$

Thus, we clearly have

$$\frac{\partial P}{\partial \lambda_j} = 0 \qquad 1 \le j \le m$$

as expected.

We now examine the jth partial derivative for $m < j \le M$. Using equation (3.5) and the fact that $\lambda_j = 0$ in the range of j under consideration, we find

$$\frac{\partial P}{\partial \lambda_j} = \left[\frac{\sum_{i=1}^{m}\lambda_i}{\sum_{i=1}^{m}\lambda_i T_i}\right]^2 \left[\frac{1}{\mu C_1} \cdot \frac{1}{(1-\rho_1)^2} - T_j\right]$$

Now (as $\lambda_j = 0$),

$$T_j = \frac{1}{\mu C_j} = \frac{1}{\mu C_1} \cdot \frac{C_1}{C_j} = \frac{1}{\mu C_1} \cdot \frac{1}{S_j^2}$$

and so

$$\frac{\partial P}{\partial \lambda_j} = \left[\frac{\sum_{i=1}^{m}\lambda_i}{\sum_{i=1}^{m}\lambda_i T_i}\right]^2 \left[\frac{1}{\mu C_1} \cdot \frac{1}{(1-\rho_1)^2} - \frac{1}{\mu C_1} \cdot \frac{1}{S_j^2}\right]$$

Therefore,

$$\frac{\partial P}{\partial \lambda_j} = \frac{1}{\mu C_1} \cdot \frac{1}{(1-\rho_1)^2} \left[\frac{\sum_{i=1}^{m} \lambda_i}{\sum_{i=1}^{m} \lambda_i T_i} \right]^2 \left[1 - \left(\frac{1-\rho_1}{S_j} \right)^2 \right]$$

or, using $\rho_1 = \rho_1(v_m) = 1 - v_m$,

$$\frac{\partial P}{\partial \lambda_j} = \frac{1}{\mu C_1} \cdot \frac{1}{(v_m)^2} \left[\frac{\sum_{i=1}^{m} \lambda_i}{\sum_{i=1}^{m} \lambda_i T_i} \right]^2 \left[1 - \left(\frac{v_m}{S_j} \right)^2 \right]$$

for $m < j \leq M$. Thus we have

$$\frac{\partial P}{\partial \lambda_j} = \begin{cases} 0 & 1 \leq j \leq m \\ K \left[1 - \left(\frac{v_m}{S_j} \right)^2 \right] & m < j \leq M \end{cases} \tag{3.40}$$

where $K > 0$.

We will now use the above evaluation of the jth partial derivative to find conditions on the given parameters S_i which insure that the solution corresponding to the root v_m is not the global optimum. To this end, we evaluate $\nabla P \cdot \mathbf{d}$ at the point given by v_m where $\mathbf{d}$ is any feasible direction. Such a feasible direction must be of the form $\mathbf{d} = (d_1, \ldots, d_M)$ where $d_j > 0$ for $m < j \leq M$. We have

$$\nabla P \cdot \mathbf{d} = \sum_{j=1}^{M} \frac{\partial P}{\partial \lambda_j} \cdot d_j$$

which, using equation (3.40), becomes

$$\nabla P \cdot \mathbf{d} = K \sum_{j=m+1}^{M} d_j \left[1 - \left(\frac{v_m}{S_j} \right)^2 \right]$$

One condition that insures $\nabla P \cdot \mathbf{d} > 0$ (and, therefore, that the solution corresponding to v_m is not optimal) is to have

$$1 - \left(\frac{v_m}{S_j} \right)^2 > 0 \qquad m < j \leq M$$

or

$$v_m < S_j \qquad m < j \leq M$$

Since $S_M \leq S_j$ for all $1 \leq j \leq M$, this condition becomes

$$v_m < S_M$$

Therefore, if the root v_m satisfies $v_m < S_M$, then the solution point corresponding to v_m (in the original M-dimensional space) is not even a local maximum, because there are feasible directions $\mathbf{d}$ such that $\nabla P \cdot \mathbf{d} > 0$ at that point.

Note that the verification of the condition $v_m < S_M$ involves the calculation of the root v_m, which is precisely what we are trying to avoid. Thus we wish to find such a condition which depends on the input parameters S_i only. Recall from equation (3.14) that we have $\rho_1 \geq 1/2$ for a critical point $(\rho_1, \ldots, \rho_m)$ of P^m. Then we must have $v_m \leq 1/2$ since $v_m = 1 - \rho_1$. Hence, for those systems for which $S_M > 1/2$, we have $v_m < S_M$, and thus, by the above argument, the solution point corresponding to v_m is not optimal. Since this is true for any $m < M$, the point corresponding to the root v_M must be globally optimal for such systems. We have proved the following

Theorem 3.5

If $S_M > 1/2$ (i.e., $C_1/C_M < 4$), then the global maximum of P for the M channel parallel network is given by the point which corresponds to the root v_M.

This yields the promised condition based on the given parameters S_i which guarantees that only one solution point (that corresponding to v_M) needs to be calculated in determining the optimal power point. Note that the result for the two channel example in the range $1/2 < S \leq 1$ follows from the above theorem. Note also that, from Theorem 3.5, if $S_M > 1/2$, then the root v_M must be real and yield a feasible solution point (a result that is difficult to prove directly).

Now let us return to the problem of finding simplifying conditions for the general case. The procedure which led to the proof of Theorem 3.5 may be generalized in the following way. Define n as the largest index $(1 \leq n \leq M)$ for which $S_n > 1/2$. Therefore, $S_i > 1/2$ (i.e., $C_i/C_1 > 1/4$) for $1 \leq i \leq n$, and $S_i \leq 1/2$ (i.e., $C_i/C_1 \leq 1/4$) for $i > n$. Note that Theorem 3.5 is the case $n = M$. Thus the ratio of the capacity of the fastest channel to that of any channel up to and including channel n is less than four ($C_1/C_i < 4$ for $1 \leq i \leq n$), while its ratio to the capacities of all channels strictly slower than channel n is at least four ($C_1/C_i \geq 4$ for $n < i$). Applying the previous argument to the subnetwork with n channels, we see that the solutions corresponding to roots $v_{n-1}, \ldots, v_1$ must yield lower power than the solution corresponding to v_n. Therefore, the optimal power point must be given by one of the roots $v_M, \ldots, v_n$, and we may ignore the others. We have shown the following

Theorem 3.6

Let n be the largest index ($1 \leq n \leq M$) such that $S_n > 1/2$. The global maximum of P for the M channel parallel network is given by a point corresponding to one of the roots $v_M, \ldots, v_n$.

This enables us to greatly simplify the determination of the optimal power point for certain networks. In terms of the procedure given at the end of section 3.6, we may insert the following new step between steps 4 and 5:

Step 4.5: Check if $S_m > 1/2$; if yes, go to Step 6

Thus as soon as $S_m > 1/2$, we may immediately disregard any subsequent roots. That is, we have found a condition which is easy to check and which (if true) enables us to eliminate the tail of our optimization procedure.

3.7.1 A Non-Reducible Example

Although, as shown in Theorems 3.5 and 3.6, we may sometimes simplify the determination of the optimal power point, the next example illustrates that such simplification is not always possible. This sample network requires that all M roots must be explicitly evaluated. The net has 65 channels ($M = 65$) with the fastest channel (number 1) having capacity $16C$ and all 64 other channels (numbered $2, \ldots, 65$) having capacity C. Thus we have $S_1 = 1$ and $S_i = 1/4$ for $2 \leq i \leq 65$. Hence, $\sum_{i=1}^{M} S_i = 17$ and $\sum_{i=1}^{M} S_i^2 = 5$. Therefore, we have

$$D_M = 9\left(\sum_{i=1}^{M} S_i\right)^2 - 8M\sum_{i=1}^{M} S_i^2 = 9(17)^2 - 8(65)(5) = 1$$

Thus

$$v_M = \frac{3(17) - 1}{4(65)} = \frac{50}{260} = \frac{5}{26}$$

and

$$u_M = \frac{3(17) + 1}{4(65)} = \frac{52}{260} = \frac{1}{5}$$

Since $0 < v_M < u_M < S_M$, this gives an example where there are two critical points, both of which are feasible interior points. Hence, any algorithm which attempts to find critical points may converge to a point which is not the global optimum (and which may not even be a local optimum). However, we know from arguments above that v_M must yield the solution with higher power. Evaluating P from v_M by equation (3.33) gives the value

$$P(v_M) \cong (.1281)(\mu C_1)^2$$

Let us now consider the subnetwork obtained by dropping the slowest channel (channel number 65). This subnetwork has 64 channels with $S_1 = 1$ and $S_i = 1/4$ for $2 \leq i \leq 64$. Thus, we have $\sum_{i=1}^{M-1} S_i = \frac{67}{4}$ and $\sum_{i=1}^{M-1} S_i^2 = \frac{79}{16}$. Therefore

$$D_{M-1} = 9\left(\sum_{i=1}^{M-1} S_i\right)^2 - 8(M-1)\sum_{i=1}^{M-1} S_i^2 = 9\left(\frac{67}{4}\right)^2 - 8(64)\left(\frac{79}{16}\right) = -\frac{47}{16}$$

Thus, the new roots v_{M-1} and u_{M-1} do not lead to a feasible solution point; these roots are not even real! It can be shown that the subnetworks with 2 through 64 channels all yield roots which are not real. The subnetwork with a single channel, of course, yields the feasible solution $\rho_1 = 1/2$ and $\rho_i = 0$ for $2 \leq i \leq 65$. This point has power value

$$P(v_1) = (\mu C_1)^2/4 > P(v_M)$$

and it is therefore the global maximum.

We first note that this gives another example of an unfair optimal operating point. In fact, 64 of the 65 users have zero throughput at this optimal point. We next observe that $M^* = 1$ while $M = 65$; therefore, as was suggested earlier, the bound given by equation (3.20) may be quite bad. We have found an example of a net such that the original network has an interior critical point (which is not globally optimal) and the only other subnetwork which has a feasible solution point is the final one (and this point is the globally optimal point). Thus, all M subnetworks must be considered in the determination of the global optimum. This network can be perturbed slightly to give examples where the original net and more than one subnet give feasible roots, but other subnets do not (simply start with more than 64 slow channels in addition to the one fast channel).

3.8 Fairness

We now demonstrate that Theorems 3.5 and 3.6 yield results concerning Jaffe's notion of fairness. We first consider Theorem 3.5. Note that the optimal point (given by v_M) is an interior point of the feasible region since $v_M \leq 1/2 < S_M$, and so it yields a fair solution point. We have shown the following

Theorem 3.7

If $S_M > 1/2$ (i.e., $C_1/C_M < 4$), then the global maximum of P for the M channel parallel network is a fair operating point.

Thus we see that if the capacity of the fastest channel is less than four times the capacity of the slowest channel, the optimal operating point with respect to power gives each channel (or user) non-zero throughput. Therefore, an unfair optimal solution can only occur when the ratio of the fastest channel capacity to the slowest is at least four. It is important to try to characterize

optimal solutions which yield channels with zero throughput because such a parallel configuration can then be collapsed to a net with fewer channels which is equivalent in terms of power. This may be valuable in the study of the power problem for general network topologies. However, as we shall see in section 3.8.1, the "width of four" property does not characterize unfair solutions. That is, there are parallel networks satisfying $C_1/C_M \geq 4$ which yield fair global maxima (i.e., each user has non-zero throughput).

The above result that $S_M > 1/2$ implies the optimal solution is fair (which used Theorem 3.5) may be generalized using Theorem 3.6 and the following claim.

Claim: If $1 \leq i \leq j \leq M$ and $S_i > 1/2$, then the solution corresponding to root v_j (for the subnetwork with j channels) is fair for user i.

We prove this claim by first noting that $v_j \leq 1/2 < S_i$, and also (since $i \leq j$, and thus the subnetwork with j channels necessarily includes channel i) that $\rho_i(v_j) = 1 - v_j/S_i$. Therefore $\rho_i > 0$, and the solution is fair for user i. Thus the claim holds, which enables us to prove the following

Theorem 3.8

If for some m $(1 \leq m \leq M)$ $S_m > 1/2$ (i.e., $C_1/C_m < 4$), then the global maximum of P for the M channel parallel network is fair for user m in the sense that channel m is assigned non-zero throughput at the optimal power point.

To prove this, we first define n as in Theorem 3.6; that is, we let n be the largest index $(1 \leq n \leq M)$ such that $S_n > 1/2$. Thus we must have $1 \leq m \leq n$. From Theorem 3.6, we know that the optimal power point corresponds to one of the roots $v_M, \ldots, v_n$. We now note that this optimal point will be fair for users 1 through n (and thus for user m) no matter which of the roots $v_M, \ldots, v_n$ yields highest power. This is true from the claim proved above, because if we let v_j yield the optimal point and i be one of these first n users, then $i \leq n \leq j$. Thus root v_j yields a point which gives non-zero throughput to each of the channels (users) $1, \ldots, n$ and therefore to user m, which proves Theorem 3.8.

This shows that if the ratio of the capacity of the fastest channel to the capacity of channel m $(1 \leq m \leq M)$ is less than four, then the optimal power point is fair for user m. We have found a condition (based on the given channel capacities) which is helpful in determining the set of users for which the optimal power point is fair, but which unfortunately does not fully characterize fairness as we shall see in section 3.8.1.

3.8.1 Fairness Characterization Counterexample

We now give the promised counterexample to the "width of four" conjecture. For the case of two channels, Theorem 3.4 shows that the optimal solution when power P is maximized is fair if and only if $C_1/C_2 < 4$. Thus Theorem 3.4 gives a characterization of fairness for $M = 2$ in terms of the given channel capacities. For the case of M channels, Theorem 3.5 shows that the optimal solution is fair if the condition $C_1/C_M < 4$ on the channel capacities holds. We now show that this property does not characterize fairness for this general case. That is, if the condition $C_1/C_M \geq 4$ holds, then it is not necessarily the case that the optimal solution is unfair. Specifically, we prove the following

Theorem 3.9

For any real number $\alpha \geq 4$, there is an M/M/1 parallel network with $C_1/C_M = \alpha$, but whose optimal solution for power P is fair.

In fact, the example network we choose is of the form $C_1/C_i = \alpha$ for $2 \leq i \leq M$. For notational convenience, define $\beta \triangleq \sqrt{\alpha}$ so that $\beta^2 = \alpha$. Therefore, we have $\beta \geq 2$. We consider a parallel network of M channels with $C_1 = \alpha C = \beta^2 C$ and $C_2 = \cdots = C_M = C$. Thus the ratio of the capacity of the single fast channel to the capacity of any of the slow channels is $\alpha = \beta^2$. We have $S_1 = 1$ and $S_2 = \cdots = S_M = 1/\beta$. We will show that there is a value of $M = M'$ large enough such that the resulting parallel system yields a fair global solution point when power P is maximized. Intuitively, this comes about since the accumulated throughput for a large number of slow channels is great enough to overcome the additional delay introduced by those slow channels.

To prove this theorem, we will show that

(i) v_M is real for large M

(ii) v_M yields a feasible interior solution point for large M

(iii) $\lim_{M \to \infty} P^M(v_M) = \infty$

Assuming (i), (ii) and (iii), we now prove Theorem 3.9. We can find an integer M' such that $v_{M'}$ is real and yields a feasible interior solution point (from (i) and (ii)), and also such that

$$P^{M'}(v_{M'}) > P^j(v_j)$$

for all $j < M'$ (from (iii)). Thus, as $P^j(v_j) = P^{M'}(v_j)$, we also have

$$P^{M'}(v_{M'}) > P^{M'}(v_j)$$

for $j < M'$. The optimal boundary point is given by v_j for some $j < M'$ (recall, that the

subnetworks are obtained by dropping the slowest channel recursively), and thus the above choice of M' shows that this critical point yields higher power than the optimal boundary point. Therefore, the solution corresponding to $v_{M'}$ must be globally optimal. From (i) and (ii), the root $v_{M'}$ is real and yields a feasible solution point which is an interior critical point of $P = P^{M'}$. Since this point is interior to the feasible region, it yields a *fair* solution, i.e., each user has non-zero throughput. This proves Theorem 3.9.

Before proving (i), (ii) and (iii), we establish some results which will assist us in these proofs. We first set $L \triangleq M - 1$, so that the network has 1 fast channel of capacity αC and L slow channels each of capacity C. We observe that $\sum_{i=1}^{M} S_i = 1 + \frac{L}{\beta}$ and $\sum_{i=1}^{M} S_i^2 = 1 + \frac{L}{\beta^2}$. The roots $u = u_M = u_{L+1}$ and $v = v_M = v_{L+1}$ are given in terms of the discriminant

$$D(L) = 9\left(\sum_{i=1}^{M} S_i\right)^2 - 8(1+L)\sum_{i=1}^{M} S_i^2 \tag{3.41}$$

Using the values for S_i, we have

$$D(L) = 9\left(1 + \frac{L}{\beta}\right)^2 - 8(1+L)\left(1 + \frac{L}{\beta^2}\right)$$

or

$$D(L) = 9\left(1 + \frac{2L}{\beta} + \frac{L^2}{\beta^2}\right) - 8\left(1 + L + \frac{L}{\beta^2} + \frac{L^2}{\beta^2}\right)$$

Multiplying and collecting terms in powers of L, we obtain

$$D(L) = 1 - \left(8 - \frac{18}{\beta} + \frac{8}{\beta^2}\right)L + \frac{L^2}{\beta^2}$$

which finally yields

$$D(L) = \frac{1}{\beta^2}\left[L^2 - (8\beta^2 - 18\beta + 8)\,L + \beta^2\right]$$

We may rewrite this expression in the form

$$D(L) = \frac{1}{\beta^2}\left[L^2 - f(\beta)\,L + \beta^2\right] \tag{3.42}$$

where we have defined

$$f(\beta) \triangleq 8\beta^2 - 18\beta + 8 \tag{3.43}$$

By differentiating with respect to β, we observe that $f(\beta)$ is strictly convex with a global minimum at $\beta = 9/8$. Thus $f(\beta)$ is strictly convex increasing for $\beta \geq 2$. Also note that $f(\beta) \geq f(2) = 4 > 0$ for $\beta \geq 2$, and so $f(\beta)$ is positive for the given (fixed) value of β of Theorem 3.9.

Writing equations (3.27) and (3.28) in terms of the parameters of the example network(s) we are studying, we have

$$u = \frac{3\left(1+\frac{L}{\beta}\right)+\sqrt{D(L)}}{4(1+L)} \tag{3.44}$$

and

$$v = \frac{3\left(1+\frac{L}{\beta}\right)-\sqrt{D(L)}}{4(1+L)} \tag{3.45}$$

We know that the solution point corresponding to v gives higher power, and thus we wish to evaluate $P(v)$ as given by equation (3.33). We first calculate

$$2u - v = \frac{3[\beta+L+\beta\sqrt{D(L)}]}{4\beta(1+L)}$$

Using equation (3.33), we have

$$P(v) = (\mu C_1)^2 v^3 \frac{[\beta+L+\beta\sqrt{D(L)}]}{\beta}$$

Since $C_1 = \beta^2 C$, we obtain

$$P(v) = (\mu C)^2 \beta^3 v^3 [\beta+L+\beta\sqrt{D(L)}] \tag{3.46}$$

We now prove (i). From equation (3.29), we see that the roots u and v are real if and only if $D(L) \geq 0$. From equation (3.42), we observe that $D(L)$ is positive for large L, and thus (i) holds.

We next prove (ii). We wish to show that, for large L, the root $v = v_M = v_{L+1}$ yields a feasible solution which is also an interior point of the feasible region. That is, we claim that $v < S_M$ for large L. This is true if and only if

$$v < \frac{1}{\beta} \tag{3.47}$$

for large L, which we now proceed to prove. Recall, from the proof of (i), that the roots u and v are real for large L. Using equation (3.45), we see that equation (3.47) is equivalent to

$$\frac{3\left(1+\frac{L}{\beta}\right)-\sqrt{D(L)}}{4(1+L)} < \frac{1}{\beta}$$

or

$$3(\beta+L) - \beta\sqrt{D(L)} < 4(1+L)$$

This last inequality yields

$$3\beta < L + 4 + \beta\sqrt{D(L)}$$

which is clearly true for large L, since $D(L)$ is positive for large L and β is fixed. This proves equation (3.47) and therefore (ii). Thus, for large L, we have shown that the root v is real and yields a feasible solution point which is an interior point of the feasible region. Therefore, it yields a *fair* solution.

We now prove (iii). We first find a lower bound on $v = v_{L+1} = v_M$ which is useful in bounding the corresponding value of power $P(v)$. We claim that

$$\frac{1}{2\beta} < v \tag{3.48}$$

for large L. Recall that, for large L, v is real and yields a feasible solution which is fair. Using equation (3.45), we see that equation (3.48) is true if and only if

$$\frac{1}{2\beta} < \frac{3(1+\frac{L}{\beta}) - \sqrt{D(L)}}{4(1+L)}$$

or

$$2(1+L) < 3(\beta+L) - \beta\sqrt{D(L)}$$

This is equivalent to

$$\beta\sqrt{D(L)} < L + 3\beta - 2 \tag{3.49}$$

Note that, since $\beta \geq 2$, we have $3\beta - 2 > 0$. Thus for equation (3.49) to hold, we need only show that

$$\beta\sqrt{D(L)} \leq L \tag{3.50}$$

for large L. Since $D(L)$ is positive for large L, we may square equation (3.50) and preserve the inequality to obtain

$$\beta^2 D(L) \leq L^2$$

Using equation (3.42), we have

$$L^2 - f(\beta)L + \beta^2 \leq L^2$$

Rearranging terms, we need only prove

$$\beta^2 \leq f(\beta)L$$

for large L. Since $f(\beta)$ is positive for $\beta \geq 2$, this inequality clearly holds for large L (recall that β is fixed). This proves equation (3.48).

From the above results, for large L (i.e., for large M), the root $v = v_{L+1} = v_M$ is real, yields a feasible solution point which is fair, and also satisfies equation (3.48). In particular, we have shown for large L that

$$\frac{1}{2\beta} < v < \frac{1}{\beta} \tag{3.51}$$

We now bound the value $P(v)$ for large L. Recall that equation (3.46) gives $P(v)$ as

$$P(v) = (\mu C)^2 \beta^3 v^3 [\beta + L + \beta\sqrt{D(L)}]$$

We use the bound given in equation (3.48) to obtain

$$P(v) > (\mu C)^2 \beta^3 \cdot \frac{1}{(2\beta)^3} [\beta + L + \beta\sqrt{D(L)}]$$

or

$$P(v) > (\mu C)^2 \frac{[\beta + L + \beta\sqrt{D(L)}]}{8} \tag{3.52}$$

for large L. Since $D(L)$ is positive for large L, the right-hand side of the above inequality increases without bound as $L \to \infty$ (i.e., as $M \to \infty$). Therefore, we have

$$\lim_{M \to \infty} P^M(v_M) = \infty \tag{3.53}$$

where P^M is the power function of an M channel parallel network. This proves (iii), and thus Theorem 3.9.

3.9 Non-Concavity of Power

Our final example network demonstrates that power need not be concave with respect to the individual throughputs. The example yields a network with an "undesirable" power function in the following sense. As mentioned above, an algorithm using only local information will not usually yield the globally optimal power point. An algorithm which makes use of global information to optimize power and which expects the function to have "nice" properties (for example, concavity or its relatives) may also fail if the power function is not well-behaved. In a recent paper [Bhar81], Jaffe and Bharath-Kumar give an example of a computer network which has a power function that is not concave. The parallel network to be introduced next furnishes another (simpler) example which has a non-concave power function. Thus the problem of optimizing power for an M/M/1 parallel network is, in general, not an instance of the well-known convex programming problem. Any optimization algorithm which makes use of the concavity of the objective function will not necessarily converge to a global (or even local) maximum if indeed the objective function is not concave.

Before introducing this example, we first calculate the Hessian matrix of second partials for the power function of an arbitrary M channel M/M/1 parallel network. We differentiate the expression for the jth partial derivative of P given in equation (3.2) in order to determine the second partials of the power function. A laborious but straightforward computation yields

$$\frac{\partial^2 P}{\partial \lambda_j^2} = \frac{2\left[\left(\sum_{i=1}^{M} \lambda_i T_i\right) - \left(\sum_{i=1}^{M} \lambda_i\right)\left(\lambda_j \frac{dT_j}{d\lambda_j} + T_j\right)\right]^2 - \left(\sum_{i=1}^{M} \lambda_i T_i\right)\left(\sum_{i=1}^{M} \lambda_i\right)^2 \left(\lambda_j \frac{d^2 T_j}{d\lambda_j^2} + 2\frac{dT_j}{d\lambda_j}\right)}{\left(\sum_{i=1}^{M} \lambda_i T_i\right)^3} \tag{3.54}$$

for $1 \le j \le M$. Similarly, the cross partials are found to be

$$\frac{\partial^2 P}{\partial \lambda_k \partial \lambda_j} = \frac{\partial^2 P}{\partial \lambda_j \partial \lambda_k} = \frac{2\left(\sum_{i=1}^{M} \lambda_i T_i\right)^2 + 2\left(\sum_{i=1}^{M} \lambda_i\right)^2 \left(\lambda_k \frac{dT_k}{d\lambda_k} + T_k\right)\left(\lambda_j \frac{dT_j}{d\lambda_j} + T_j\right)}{\left(\sum_{i=1}^{M} \lambda_i T_i\right)^3}$$

$$- \frac{2\left(\sum_{i=1}^{M} \lambda_i T_i\right)\left(\sum_{i=1}^{M} \lambda_i\right)\left[\left(\lambda_k \frac{dT_k}{d\lambda_k} + T_k\right) + \left(\lambda_j \frac{dT_j}{d\lambda_j} + T_j\right)\right]}{\left(\sum_{i=1}^{M} \lambda_i T_i\right)^3} \tag{3.55}$$

for $1 \le j \ne k \le M$.

Let us evaluate these second partials at critical points of the power function P. At such points equation (3.3) holds, and so equation (3.54) simplifies to

$$\frac{\partial^2 P}{\partial \lambda_j^2} = \frac{2\left(\sum_{i=1}^{M} \lambda_i T_i\right) - \left(\sum_{i=1}^{M} \lambda_i\right)^2 \left(\lambda_j \frac{d^2 T_j}{d\lambda_j^2} + 2\frac{dT_j}{d\lambda_j}\right)}{\left(\sum_{i=1}^{M} \lambda_i T_i\right)^2}$$

We recall, for M/M/1,

$$\frac{dT_j}{d\lambda_j} = T_j^2$$

so that

$$\frac{d^2 T_j}{d\lambda_j^2} = 2T_j \frac{dT_j}{d\lambda_j} = 2T_j^3$$

Therefore, we find that

$$\lambda_j \frac{d^2 T_j}{d\lambda_j^2} + 2\frac{dT_j}{d\lambda_j} = 2T_j\left(\lambda_j \frac{dT_j}{d\lambda_j} + T_j\right)$$

Using this in our expression for the second partial plus equation (3.3) once again yields

$$\frac{\partial^2 P}{\partial \lambda_j^{\,2}} = \frac{2\left(\sum_{i=1}^{M} \lambda_i T_i\right) - 4T_j\left(\sum_{i=1}^{M} \lambda_i\right)\left(\sum_{i=1}^{M} \lambda_i T_i\right)}{\left(\sum_{i=1}^{M} \lambda_i T_i\right)^2}$$

Simplifying this result and utilizing equation (3.25), we finally have

$$\frac{\partial^2 P}{\partial \lambda_j^{\,2}} = \frac{2}{\sum_{i=1}^{M} \lambda_i T_i}\left[1 - \frac{2T_1}{S_j}\sum_{i=1}^{M} \lambda_i\right] \tag{3.56}$$

In a similar manner, if we evaluate the cross partials at critical points of P, equation (3.55) becomes

$$\frac{\partial^2 P}{\partial \lambda_k \partial \lambda_j} = \frac{\partial^2 P}{\partial \lambda_j \partial \lambda_k} = \frac{2}{\sum_{i=1}^{M} \lambda_i T_i} \tag{3.57}$$

Thus, the Hessian matrix $\mathbf{H}$ evaluated at points where $\nabla P = 0$ is given by

$$\mathbf{H} = \frac{2}{\sum_{i=1}^{M} \lambda_i T_i}\begin{bmatrix} 1 - 2T_1\sum_{i=1}^{M} \lambda_i & 1 & \cdots & 1 \\ 1 & 1 - \frac{2T_1}{S_2}\sum_{i=1}^{M} \lambda_i & \cdots & 1 \\ \vdots & \vdots & & \vdots \\ 1 & 1 & \cdots & 1 - \frac{2T_1}{S_M}\sum_{i=1}^{M} \lambda_i \end{bmatrix} \tag{3.58}$$

We now introduce our final example. This M/M/1 parallel network has one fast channel of capacity $4C$ and three slow channels, each of capacity C, so that $M = 4$. We observe that $S_1 = 1$ and $S_2 = S_3 = S_4 = 1/2$. Hence, $\sum_{i=1}^{M} S_i = \frac{5}{2}$ and $\sum_{i=1}^{M} S_i^2 = \frac{7}{4}$. We therefore have

$$D = 9\left(\sum_{i=1}^{M} S_i\right)^2 - 8M\sum_{i=1}^{M} S_i^2 = 9\left(\frac{5}{2}\right)^2 - 8(4)\left(\frac{7}{4}\right) = \frac{1}{4}$$

The root $v = v_M$ is

$$v = \frac{3\left(\frac{5}{2}\right) - \frac{1}{2}}{4(4)} = \frac{7}{16}$$

while $u = u_M$ is

$$u = \frac{3(\frac{5}{2}) + \frac{1}{2}}{4(4)} = \frac{1}{2}$$

The solution corresponding to v is $\rho_1 = 9/16$ and $\rho_2 = \rho_3 = \rho_4 = 1/8$. It can be shown that this is the global maximizer of P (all subnetworks yield the same root of 1/2). We see that this point is an interior point so that every user has non-zero throughput (i.e., the global optimum is fair) even though the ratio of the capacity of the fastest channel to that of the slowest is equal to four. This is unlike the situation for two channels and shows that the "width of four" property does not characterize unfair solutions, even in the case of $M = 4$ channels.

Another interesting property of this net is that its power function is not concave. First recall the well-known result [Marl78] that a function $f: S \to R$, where S is an open set and $S \subseteq R^n$, is concave if and only if its Hessian matrix of second partials is negative semidefinite at all points of the domain of f. Thus, to show that P is not concave, we need only exhibit a point such that the Hessian is not negative semidefinite there. Next recall that a necessary (but *not* sufficient) condition for a matrix to be negative semidefinite is that the determinants of the leading principal submatrices, proceeding in order down the main diagonal, are alternately ≤ 0 and ≥ 0, with the first being nonpositive. Thus, a counterexample to the concavity of P may be found by exhibiting a point at which the Hessian of P does not satisfy the above condition. This we now proceed to do.

Consider the point corresponding to the root u. This point has coordinates $\rho_1 = 1/2$ and $\rho_2 = \rho_3 = \rho_4 = 0$ and is a critical point of P. Thus we may evaluate the Hessian $\mathbf{H}$ at this point using equation (3.58). We have

$$\mathbf{H} = 2\begin{bmatrix} -1 & 1 & 1 & 1 \\ 1 & -3 & 1 & 1 \\ 1 & 1 & -3 & 1 \\ 1 & 1 & 1 & -3 \end{bmatrix}$$

The four leading principal submatrices are

$$[-2], \quad \begin{bmatrix} -2 & 2 \\ 2 & -6 \end{bmatrix}, \quad \begin{bmatrix} -2 & 2 & 2 \\ 2 & -6 & 2 \\ 2 & 2 & -6 \end{bmatrix}, \quad \text{and } \mathbf{H}$$

with corresponding determinants $-2, 8, 0$, and -512. These determinants do not satisfy the previously mentioned condition, and so $\mathbf{H}$ is not negative semidefinite at the point $(1/2, 0, 0, 0)$.

(Although this point is on the boundary of the feasible region, we note that, by the continuity of the entries of the Hessian, there are interior points at which the Hessian is also not negative semidefinite.) Therefore, we have the promised example of a parallel network such that its power function is not concave. Thus any algorithm which uses concavity to find a global maximum of a function would, in general, fail for this power function P.

This concludes our analysis of power for the M/M/1 parallel network. We have found a solution procedure for the general case of M channels, and we have derived the exact analytical solution for the two channel net. The solution of this optimization problem was not an easy task, however. Examples of parallel nets with unfair optimal operating points, with the local power point differing from the global power point, and also with a nonconcave power function were presented. The difficulty of the method of solution and the various "bad" examples which resulted suggest that the power problem for general computer network configurations may perhaps be hard to solve and yield undesirable operating points. In the next chapter we extend the analysis in two different ways, which also allow us to avoid these undesirable operating points in many cases. In the first part of chapter 4, we restrict our attention to generalized network problem statements. We find that we are able to solve some of these problems and that they yield intuitively pleasing results. In the second part of chapter 4, we shall slightly alter the definition of power (still preserving the idea of a tradeoff function) and obtain several power functions, one of which (first introduced by Kleinrock) has very pleasing qualities. We also compare the various closely related notions of power which have appeared in the literature.

CHAPTER 4
Extensions of the Power Problem

In the previous two chapters, the problem of optimizing power was studied for networks with simple topologies. For example, series networks were examined in chapter 2 while parallel networks were studied in both chapters 2 and 3. In the first part of this chapter, we extend our analysis of power to general network topologies. We introduce several optimization problem formulations and indicate both the appropriate physical situations they are meant to model and the level of difficulty of solution for each representation. Thus we alter the constraints of the optimization problem and/or the decision variables over which we optimize. In so doing, we avoid many of the undesirable properties found in chapter 3. In the second part of this chapter we extend our analysis by considering other objective functions in our optimization problem; that is, we study other definitions of power. We find that one such definition, first introduced by Kleinrock, yields optimal solution points which no longer possess many of the negative characteristics which appeared as part of our parallel network analysis in chapter 3.

4.1 The Power Problem for General Network Topologies

We consider a network of N nodes (switching computers) and M channels of arbitrary topology. Our notation expands upon that of chapter 1 and again follows volume II of Kleinrock [Klei76]. We assume that each source-destination pair (j,k) of nodes of the network represents a potential user (or users) who wishes to send messages from (a HOST connected to) node j to (a HOST connected to) node k. We assume that the traffic originating from this user is Poisson with a rate of γ_{jk} messages per second. The total throughput of the system is then

$$\gamma = \sum_{j=1}^{N}\sum_{k=1}^{N}\gamma_{jk} \tag{4.1}$$

As before we let C_i be the capacity of channel i and let λ_i be the rate of traffic on the ith channel in messages per second. Note that λ_i is the sum (over all (j,k) pairs) of the portion of γ_{jk} traffic which uses channel i. That is, if we define p_{ijk} as the fraction of γ_{jk} which traverses channel i ($p_{ijk}=0$ when no (j,k) traffic uses channel i), then

$$\lambda_i = \sum_{j=1}^{N}\sum_{k=1}^{N}p_{ijk}\gamma_{jk} \tag{4.2}$$

The p_{ijk} are determined from the routing scheme; fixed routing (where unique paths for each γ_{jk} are chosen) is not necessary for this analysis, but it may hold in a particular network under

study (for example, in the case of virtual circuits). Using Little's result, the average delay T of a message in its journey through the network may be determined in terms of the mean channel delays T_i. As shown in [Klei76], since

$$\bar{N} = \sum_{i=1}^{M} \bar{N}_i$$

by Little's result we have

$$\gamma T = \sum_{i=1}^{M} \lambda_i T_i$$

and so

$$T = \sum_{i=1}^{M} \frac{\lambda_i}{\gamma} T_i \tag{4.3}$$

Recall that $T_i = \bar{x}_i/(1-\rho_i)$ for M/M/1, where $\rho_i = \lambda_i \bar{x}_i$, $\bar{x}_i = \bar{b}/C_i$ and the mean packet length is $\bar{b} = 1/\mu$. Thus $T_i = 1/(\mu C_i - \lambda_i)$, and we have

$$T = \sum_{i=1}^{M} \frac{\lambda_i}{\gamma} \left[\frac{1}{\mu C_i - \lambda_i} \right] \tag{4.4}$$

Therefore, the power function P is simply

$$P = \frac{\gamma}{T} = \frac{\gamma^2}{\sum_{i=1}^{M} \lambda_i T_i} = \frac{\left(\sum_{j=1}^{N} \sum_{k=1}^{N} \gamma_{jk} \right)^2}{\sum_{i=1}^{M} \frac{\lambda_i}{\mu C_i - \lambda_i}} \tag{4.5}$$

Finally, using equation (4.2), we may write P in terms of the traffic matrix $\{\gamma_{jk}\}$ and the routing fractions $\{p_{ijk}\}$ as

$$P = \frac{\left(\sum_{j=1}^{N} \sum_{k=1}^{N} \gamma_{jk} \right)^2}{\sum_{i=1}^{M} \left[\frac{\sum_{j=1}^{N} \sum_{k=1}^{N} p_{ijk} \gamma_{jk}}{\mu C_i - \sum_{j=1}^{N} \sum_{k=1}^{N} p_{ijk} \gamma_{jk}} \right]} \tag{4.6}$$

Before introducing an appropriate maximization problem involving power, we consider the parallel net studied in both chapters 2 and 3. The previous analyses of parallel nets yielded two different optimization problems depending upon whether the routing of messages in the net was a known quantity or an unknown variable. In chapter 2 the routing was assumed to be known (i.e. the fraction of traffic p_i on the ith channel was given). The resulting optimization problem was mathematically simple and was easily solved. In this problem the only variable to be optimized was the total throughput of the system since the routing percentages were

predetermined. Here our interest was solely in *flow control*, or how to regulate traffic rates at network entry points [Gerl80]. In chapter 3 we also had to optimize the *routing*, or how to best direct traffic from sources to destinations [Schw80]. In this case, the maximization of power was a mathematically difficult problem. Using the parallel network results as a guide, we will introduce several power problem formulations for arbitrary M/M/1 networks. The most general such optimization problem is not easily solvable, but other formulations (which model more accurately the physical situation of interest to us) can and will be solved to yield insights into the behavior of networks at the optimal power point.

4.1.1 Power Problem Formulations

We first consider the three optimization problems (CA, FA and CFA) introduced in section 1.1.2, but with the objective of minimizing delay replaced by that of maximizing power. In all three power is maximized for a given network topology and a given traffic matrix $\{\gamma_{jk}\}$. Thus in each case the throughput γ is known and, since P ($=\gamma/T$) is a constant divided by T, maximizing power is identical to minimizing total network delay. Therefore the resulting optimization problems are equivalent to the problems (CA, FA and CFA) described above in chapter 1. We will use this equivalence later in analyzing other power problems.

We now introduce several new optimization problems. In each case the objective is to maximize the power of the network, but the problems differ as to the variables over which we optimize. Constraints present in every case include bounds (based on capacity) on flow for each channel and conservation of flow for all nodes. In all the formulations given below, we always assume that the network topology is given (but arbitrary) and that the channel capacities C_i are known.

The first new formulation we introduce (power formulation 1 or PF1 for short) is the most general of the formulations. Here both the external traffic rates and their subsequent routing paths through the network are unknown. Thus we wish to find the traffic matrix $\{\gamma_{jk}\}$ and the routing fractions $\{p_{ijk}\}$ which optimize power.

PF1	Given:	Capacities $\{C_i\}$ and network topology
	Maximize:	P
	With respect to:	traffic matrix $\{\gamma_{jk}\}$ and routing $\{p_{ijk}\}$

If we assume the routing is known (i.e. the p_{ijk} are given), we may optimize over only γ_{jk} and then use equation (4.2) to find the channel flows. This yields our second formulation (called PF2) which requires more restrictive assumptions than the more general problem PF1.

PF2	Given:	Capacities $\{C_i\}$, routing $\{p_{ijk}\}$, and network topology
	Maximize:	P
	With respect to:	traffic matrix $\{\gamma_{jk}\}$

We note that if the network topology precludes the possibility of alternate routing (for example, a tree network), then PF1 reduces to PF2 and the two problems are equivalent. In this case, p_{ijk} is 1 or 0 depending on whether or not channel i is in the unique path connecting nodes j and k. We also note that the power problem for a virtual circuit network where each user's path is known is an instance of PF2.

Let us consider these formulations as applied to a parallel network topology with M channels. We first regard the net as representing a single user with multiple paths from source to destination (see Figure 4.1).

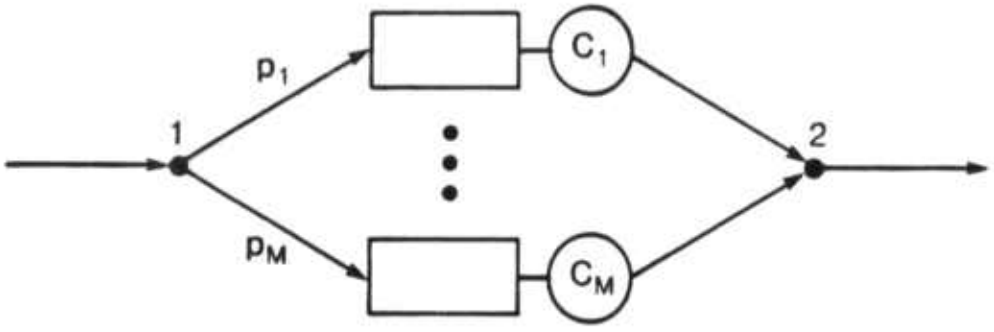

Figure 4.1 A Single User with Multiple Paths

That is, the network has two nodes, and the M channels all connect source node 1 with destination node 2. Thus $p_{i12} \triangleq p_i$ is the fraction of (total) traffic using channel i. In this case PF1 reduces to the problem studied in chapter 3, and PF2 reduces to the M/M/1 parallel net analyzed in chapter 2.

Now let us regard our parallel net as modeling a set of multiple users each with his own path (see Figure 4.2).

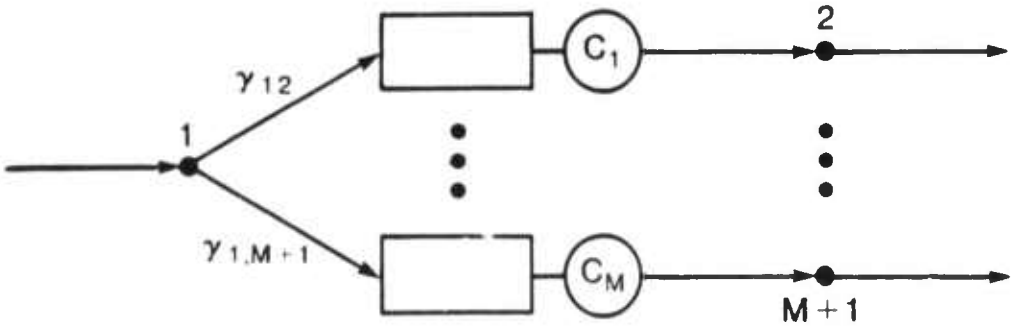

Figure 4.2 Multiple Users with Single Paths

That is, we have M users and user i sends messages over channel i. So there is one source node and M destination nodes, and channel i connects source node 1 with destination node $i+1$. In this case, the network topology is a tree, and $p_{ijk} = 1$ or 0 for all pairs (j, k) (i.e. we have fixed routing). Thus PF1 is equivalent to PF2 which is equivalent to the problem studied in chapter 3.

The above discussion of the parallel network and the results of chapter 3 show that both PF1 and PF2 may yield unfair optimal power points. These formulations may also lead to maximization problems which have an objective function which is not concave. In fact, all the negative aspects of the problem solved in chapter 3 are present in both PF1 and PF2. These formulations have led authors such as Jaffe [Jaff81] to conclude that power (as introduced by Giessler) [Gies78] is not the type of objective one wants to employ in the design and analysis of computer networks. In the spirit of these arguments, we also examine extensions of power in the next section which have more desirable mathematical properties. In the remainder of this section we take the position that it is not the objective function (power) that should be changed, but rather it is the formulation of the problem itself that should be altered. Below we reformulate the optimization problem under consideration and find that, not only is this new problem mathematically manageable, but the new formulations more adequately represent the type of real world situation we are likely to encounter and wish to model.

Consider then the problems PF1 and PF2. In both cases the traffic matrix $\{\gamma_{jk}\}$ is an unknown quantity, and the γ_{jk} are variables of optimization. The resulting solution (for example in chapter 3) may be unfair in the sense that some users are assigned zero throughput (their $\gamma_{jk} = 0$). Note that these formulations model the situation where nothing is known about the external input traffic to the net. But this may not be the case in a real situation. Various characteristics of input traffic may in fact be known. For example, many routing and flow control protocols take advantage of past history of the network. Of course this may involve the added overhead of different types of control messages.

Let us suppose that we do have some indication of the user traffic and reformulate our power problem to take advantage of this extra information. (In a decentralized environment such as the one discussed in [Jaff81], this type of global information may not be available.) We assume that we know each user's external traffic as a fraction of the total input, even though the actual traffic matrix $\{\gamma_{jk}\}$ is not given. That is, a *relative* traffic matrix $\{r_{jk}\}$ is known, where

$$r_{jk} = \frac{\gamma_{jk}}{\gamma} \tag{4.7}$$

Given $\{r_{jk}\}$, we may scale this (relative) traffic up and down. That is, for any $\alpha > 0$, scaling by α gives a traffic matrix $\{\alpha r_{jk}\}$ which we may consider as the external traffic to the network. The scaling factor α is also known as the traffic level [Gerl77]. Note that the relationship given in equation (4.7) imposes constraints on the possible traffic matrices we may consider in optimizing power, since the ratios of the individual traffic to the total throughput are given quantities.

We seek a number α^* such that the assignment of traffic $\{\alpha^* r_{jk}\}$ will optimize power over all possible choices of α. (Note that we will not be considering all positive α, because the channel capacities and network topology restrict the values of the scaling factor which yield feasible traffic matrices.) We observe that, for different values of α, different routing assignments for the corresponding αr_{jk} may be used. Thus we scale $\{r_{jk}\}$ to find the optimal power point. Once again we formulate two optimization problems depending on whether the routing is given or is an unknown variable. Our next problem (call it PF3) assumes that the routing procedure is not known.

PF3	Given:	Capacities $\{C_i\}$, network topology, and relative traffic matrix $\{r_{jk}\}$
	Maximize:	P
	With respect to:	traffic level α and routing $\{p_{ijk}\}$

If we assume the routing is known (the p_{ijk} are given), we obtain the following formulation (call it PF4).

PF4	Given:	Capacities $\{C_i\}$, routing $\{p_{ijk}\}$, network topology, and relative traffic matrix $\{r_{jk}\}$
	Maximize:	P
	With respect to:	traffic level α

As was true for our earlier formulations, we note that PF3 and PF4 are equivalent problems if no alternate routing is possible (for example, a tree network). If we assume the net has a single user, we may obtain other simplifications. In this case, the traffic matrix consists of one entry, and scaling a relative traffic matrix to maximize power is identical to optimizing over the external traffic matrix. Thus we see that in the case of a single user, PF1 and PF3 are equivalent, while PF2 and PF4 are also equivalent. We also note an important property of the optimal solution for both PF3 and PF4, namely that it is fair. Every user (source-destination pair (j,k)) is assigned a throughput $\gamma_{jk}^* = \alpha^* r_{jk}$ at the optimal solution. Since r_{jk} is positive for any such user (otherwise, by convention, he is not included in the model being studied), the throughput γ_{jk}^* is not zero. Thus the optimal solution is fair to all users. We note that in the optimal solution to PF3, some channels may have zero flow depending on the optimal routing. But still each user will have positive throughput proportional to his "needs".

We apply these formulations to the parallel network. If we regard the net as modeling a single user with multiple paths (as in Figure 4.1), PF3 is equivalent to PF1 which is the problem analyzed in chapter 3. Also PF4 is equivalent to PF2 which is the M/M/1 parallel net studied in chapter 2 (p_i corresponds to p_{i12}). If we regard the parallel net as representing multiple users each with a fixed path (as in Figure 4.2), then PF3 is equivalent to PF4 which is equivalent to the problem studied in chapter 2. In this latter case, p_i of chapter 2 corresponds to $r_{1,i+1}$ for the model of Figure 4.2.

4.1.2 Analysis of PF4

We now concentrate on the optimization problem we have designated as PF4, which is the easiest of the four to solve. We assume that a relative traffic matrix $\{r_{jk}\}$ and routing $\{p_{ijk}\}$ are given, and we wish to optimize power with respect to the traffic level α. Consider an $\alpha > 0$ which yields a feasible traffic pattern under the above assumptions. Then the external traffic satisfies $\gamma_{jk} = \alpha r_{jk}$ for all source-destination pairs (j, k). The system throughput is thus

$$\gamma = \gamma(\alpha) = \sum_{j=1}^{N}\sum_{k=1}^{N}\gamma_{jk} = \alpha\sum_{j=1}^{N}\sum_{k=1}^{N}r_{jk}$$

or, since the r_{jk} sum to 1, we have

$$\gamma = \gamma(\alpha) = \alpha \tag{4.8}$$

Thus choosing a scaling factor which yields the maximum power point is equivalent to finding the total system throughput γ which maximizes power.

Next we consider the channel flows. Let us define $\lambda_i(\alpha)$ to be the flow of messages on channel i for the traffic level α. Then equation (4.2) yields

$$\lambda_i(\alpha) = \alpha\sum_{j=1}^{N}\sum_{k=1}^{N}p_{ijk}\,r_{jk} \tag{4.9}$$

Thus we see that the channel flows are also scaled by α. Note that the double sum in equation (4.9) represents the flow on channel i for the case $\gamma = \alpha = 1$. If this double sum is zero, then no traffic is routed over channel i (for any $\alpha > 0$), and we may as well disregard this channel in our optimization. Therefore we assume that the routing and the relative traffic matrix are such that at least some traffic is routed over each of the i channels (if not, just disregard channels with zero flow and renumber the remaining ones). This assumption does not alter our problem and simplifies subsequent notation (we no longer need to remember explicitly which channels have zero traffic and which do not). Now equations (4.8) and (4.9) yield

$$\frac{\lambda_i(\alpha)}{\gamma(\alpha)} = \sum_{j=1}^{N}\sum_{k=1}^{N}p_{ijk}\,r_{jk}$$

which does not depend on α (and thus may be denoted simply as λ_i/γ) and is positive by assumption. Let us define (for all $1 \le i \le M$) the ratios

$$R_i \triangleq \frac{\lambda_i}{\gamma} \tag{4.10}$$

so that R_i is the fraction of the total throughput that uses channel i. For problem PF4, we have shown that R_i is constant (and positive) for all values of the traffic level α.

Let us optimize P with respect to α, which is identical to optimizing with respect to γ by equation (4.8). Rewriting equation (4.4) in terms of R_i, we have

$$T = \sum_{i=1}^{M} R_i \left[\frac{1}{\mu C_i - R_i \gamma} \right] \tag{4.11}$$

Since $R_i > 0$ for all $1 \leq i \leq M$ by our convention, this equation may also be written as

$$T = \sum_{i=1}^{M} \frac{1}{\mu (C_i/R_i) - \gamma} \tag{4.12}$$

From equation (4.12), we see that the values of γ $(=\alpha)$ which yield feasible traffic matrices are $0 \leq \gamma < \min_{1 \leq i \leq M} \mu (C_i/R_i)$. Figure 4.3 shows a typical (γ, T) curve from equation (4.12).

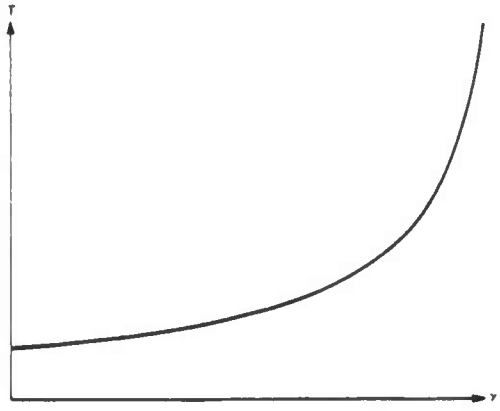

Figure 4.3 Throughput-Delay Profile for PF4

Note that $P = 0$ at the two endpoints of the feasible interval, since the throughput is zero in one case and the mean delay is infinite in the other case. Therefore, P is maximized at an interior point of this interval. By arguments from chapter 2, the derivative of P with respect to γ is zero at such a point.

We now show that power P for formulation PF4 is a strictly concave function of the traffic level γ, and therefore it has a unique maximum in the feasible region. Since $P = \gamma/T$, the first two derivatives of P with respect to γ are

$$\frac{dP}{d\gamma} = \frac{T - \gamma \frac{dT}{d\gamma}}{T^2} \tag{4.13}$$

and

$$\frac{d^2P}{d\gamma^2} = \frac{2\gamma \left(\frac{dT}{d\gamma}\right)^2 - \gamma T \frac{d^2T}{d\gamma^2} - 2T \frac{dT}{d\gamma}}{T^3} \tag{4.14}$$

From equation (4.12) we have

$$\frac{dT}{d\gamma} = \sum_{i=1}^{M} \left[\frac{1}{\mu(C_i/R_i) - \gamma} \right]^2$$

and

$$\frac{d^2T}{d\gamma^2} = 2 \sum_{i=1}^{M} \left[\frac{1}{\mu(C_i/R_i) - \gamma} \right]^3$$

Now note that

$$2T \frac{dT}{d\gamma} = 2 \left[\sum_{i=1}^{M} \frac{1}{\mu(C_i/R_i) - \gamma} \right] \left[\sum_{i=1}^{M} \left[\frac{1}{\mu(C_i/R_i) - \gamma} \right]^2 \right] > 0$$

so that from equation (4.14) we have

$$\frac{d^2P}{d\gamma^2} < \frac{2\gamma \left(\frac{dT}{d\gamma} \right)^2 - \gamma T \frac{d^2T}{d\gamma^2}}{T^3}$$

We now show that

$$2\gamma \left(\frac{dT}{d\gamma} \right)^2 - \gamma T \frac{d^2T}{d\gamma^2} \leq 0$$

The above inequality is equivalent to

$$2\gamma \left[\sum_{i=1}^{M} \left[\frac{1}{\mu(C_i/R_i) - \gamma} \right]^2 \right]^2 - 2\gamma \left[\sum_{i=1}^{M} \frac{1}{\mu(C_i/R_i) - \gamma} \right] \left[\sum_{i=1}^{M} \left[\frac{1}{\mu(C_i/R_i) - \gamma} \right]^3 \right] \leq 0$$

We observe that this holds by using the Cauchy-Schwarz inequality

$$\left(\sum_i x_i y_i \right)^2 \leq \left(\sum_i x_i^2 \right) \left(\sum_i y_i^2 \right)$$

with

$$x_i = \left[\frac{1}{\mu(C_i/R_i) - \gamma} \right]^{1/2}$$

and

$$y_i = \left[\frac{1}{\mu(C_i/R_i) - \gamma} \right]^{3/2}$$

Thus we have shown

$$\frac{d^2P}{d\gamma^2} < 0$$

and so P is a *strictly concave* function of γ. Therefore P has a *unique* maximum in the range $0 < \gamma < \min_{1 \leq i \leq M} \mu(C_i/R_i)$.

Since we are dealing with a single variable optimization problem (involving γ) which is similar in form to the problems of chapter 2, we may follow the approach developed there. The value γ^* which maximizes P satisfies the equation of Kleinrock

$$T^* = \gamma^* \left.\frac{dT}{d\gamma}\right|_{\gamma=\gamma^*} \tag{4.15}$$

which is equivalent to

$$\bar{N}^* = \gamma^* T^* = (\gamma^*)^2 \left.\frac{dT}{d\gamma}\right|_{\gamma=\gamma^*} \tag{4.16}$$

In order to find such a point, we first differentiate equation (4.11) with respect to γ and find

$$\frac{dT}{d\gamma} = \sum_{i=1}^{M} R_i \left[\frac{R_i}{(\mu C_i - R_i\gamma)^2}\right]$$

or

$$\frac{dT}{d\gamma} = \sum_{i=1}^{M} R_i^2 \left[\frac{1}{\mu C_i - \lambda_i}\right]^2$$

This we recognize as

$$\frac{dT}{d\gamma} = \sum_{i=1}^{M} R_i^2 T_i^2 \tag{4.17}$$

Multiplying equation (4.17) by γ^2 and using equation (4.16), we find that the average number of messages in the network at the optimal power point satisfies

$$\bar{N}^* = \sum_{i=1}^{M} (\lambda_i^*)^2 (T_i^*)^2$$

Using Little's result, we find (for PF4)

$$\bar{N}^* = \sum_{i=1}^{M} \bar{N}_i^* = \sum_{i=1}^{M} (\bar{N}_i^*)^2 \tag{4.18}$$

For M/M/1 we have (since $R_i > 0$)

$$\bar{N}_i = \frac{R_i\gamma}{\mu C_i - R_i\gamma} = \frac{\gamma}{\mu(C_i/R_i) - \gamma}$$

Therefore, equation (4.18) may be rewritten in terms of the unknown variable γ (which, recall, is identical to the scaling factor α) as

$$\sum_{i=1}^{M} \frac{\gamma^*}{\mu(C_i/R_i) - \gamma^*} = \sum_{i=1}^{M} \left[\frac{\gamma^*}{\mu(C_i/R_i) - \gamma^*}\right]^2 \tag{4.19}$$

Thus the optimal scaling factor is simply the root of a polynomial in γ (equivalent to equation (4.19) above) which lies within the open interval $(0, \min_{1 \le i \le M} \mu|C_i/R_i|)$. The root is unique by the strict concavity of P and may be determined by any standard root finding procedure.

Note that equation (4.18) has appeared before; it is identical to equations from chapter 2 for the M/M/1 series network and for the M/M/1 parallel network (with known routing). This is not surprising since both of the above models are particular instances of the network problem given by PF4. But equation (4.18) also appears in chapter 3 as characterizing the optimal power point for the M/M/1 parallel network with *unknown* routing (which is *not* an instance of PF4). Using this as motivation, let us try to extend equation (4.17) to our other problem formulations (PF1, PF2, and PF3). We consider the optimization problem PF1 (the most general of the four formulations). Since the power function (of that problem) is continuous and the feasible region is compact (closed and bounded), a maximal power point does exist. Thus there is a traffic matrix $\{\gamma_{jk}^*\}$ and routing $\{p_{ijk}^*\}$ (and therefore flows $\{\lambda_i^*\}$) which maximize power. This gives a relative traffic matrix $\{r_{jk}^*\}$ if we define $r_{jk}^* = \gamma_{jk}^*/\gamma^*$ for all (j,k) pairs. We let M^* be the number of channels at maximum power with non-zero flow (as in chapter 3). By renumbering if necessary, we may assume that they are channels $1 \leq i \leq M^*$. Let us consider an instance of PF4 with the above routing and relative traffic matrix as given quantities. The optimal traffic level must then be $\alpha^* = \gamma^*$, and so

$$\overline{N}^* = \sum_{i=1}^{M^*} \overline{N}_i^* = \sum_{i=1}^{M^*} (\overline{N}_i^*)^2 \tag{4.20}$$

Since $\lambda_i^* = 0$ for $i > M^*$, we see that equation (4.18) must also hold for PF1. A similar argument shows that equation (4.20) (and thus equation (4.18)) holds for the other formulations as well. We thus have the following

Theorem 4.1

For the four power problem formulations given above (PF1, PF2, PF3, and PF4), the average number in the M/M/1 network at maximum power satisfies the relationship

$$\overline{N}^* = \sum_{i=1}^{M} \overline{N}_i^* = \sum_{i=1}^{M} (\overline{N}_i^*)^2 \tag{4.21}$$

By using an argument identical to one from chapter 2, we now observe that equation (4.21) may be written in an interesting dual form. As in chapter 2, we first rewrite equation (4.21) as

$$\sum_{i=1}^{M} \overline{N}_i^* (1 - \overline{N}_i^*) = 0 \tag{4.22}$$

Defining $Q_i \triangleq 1 - \overline{N}_i^*$, we have

$$\sum_{i=1}^{M} (1 - Q_i) Q_i = 0$$

or

$$\sum_{i=1}^{M} Q_i = \sum_{i=1}^{M} (Q_i)^2$$

This proves the following

Theorem 4.2

For the four power problem formulations given above (PF1, PF2, PF3, and PF4), the average number in the M/M/1 network at maximum power satisfies the relationship

$$\sum_{i=1}^{M} (1 - \bar{N}_i^*) = \sum_{i=1}^{M} (1 - \bar{N}_i^*)^2 \tag{4.23}$$

This gives the dual equation which was mentioned above.

Using equation (4.21) in a manner identical to that of chapter 2, we may also obtain an upper bound on the average number in system at the optimal power point for any of the above four formulations. We repeat this earlier argument here for clarity. From equation (4.21) we have

$$\bar{N}^* = \sum_{i=1}^{M} \bar{N}_i^* = 2 \sum_{i=1}^{M} \bar{N}_i^* - \sum_{i=1}^{M} (\bar{N}_i^*)^2$$

Adding and subtracting the number M from the right-hand side of this equation and collecting terms in i yields

$$\bar{N}^* = M - \sum_{i=1}^{M} \left[1 - 2\bar{N}_i^* + (\bar{N}_i^*)^2 \right]$$

or

$$\bar{N}^* = M - \sum_{i=1}^{M} (1 - \bar{N}_i^*)^2$$

Thus we have the following

Theorem 4.3

For the four power problem formulations given above (PF1, PF2, PF3, and PF4), the average number in an M/M/1 network (with M channels) at maximum power satisfies

$$\bar{N}^* \leq M \tag{4.24}$$

We now observe that the above argument may be applied to equation (4.20) instead of equation (4.21). Since $M^* \leq M$, this gives the tighter bound

$$\overline{N}^* \leq M^* \tag{4.25}$$

Note that $M^* = M$ for formulation PF4 by convention (all channels were assumed to have non-zero flow). However, since the other three problems (PF1, PF2, and PF3) involve determining the routing procedure and/or the traffic matrix, in these cases the value of M^* depends on the given system parameters (such as topology and channel capacities). For such problems, the bound based on M in equation (4.24) may not be very good (see the example in section 3.7.1 of chapter 3). However, the number of channels M is a known quantity, while M^* may only be determined after the particular network under study has been optimized.

Theorem 4.1 (in the equivalent form of equation (4.22)) also enables us to determine bounds on the individual $\overline{N}_i^*$. Since $\overline{N}_i^* > 0$ for at least one index i (all i for PF4), we see that if

$$\overline{N}_i^* < 1 \qquad 1 \leq i \leq M$$

then the sum in equation (4.22) would be positive. Similarly the condition

$$\overline{N}_i^* > 1 \qquad 1 \leq i \leq M$$

cannot occur either, for then the sum would be negative. Thus we have shown that

$$\max_{1 \leq i \leq M} \overline{N}_i^* \geq 1 \geq \min_{1 \leq i \leq M} \overline{N}_i^* \tag{4.26}$$

Since $\overline{N}_i = \rho_i/(1-\rho_i)$ for M/M/1, we see that equation (4.26) gives

$$\max_{1 \leq i \leq M} \rho_i^* \geq \frac{1}{2} \geq \min_{1 \leq i \leq M} \rho_i^* \tag{4.27}$$

In an identical manner equation (4.20) yields

$$\max_{1 \leq i \leq M^*} \overline{N}_i^* \geq 1 \geq \min_{1 \leq i \leq M^*} \overline{N}_i^* \tag{4.28}$$

and thus also

$$\max_{1 \leq i \leq M^*} \rho_i^* \geq \frac{1}{2} \geq \min_{1 \leq i \leq M^*} \rho_i^* \tag{4.29}$$

Note that the upper bounds in equations (4.28) and (4.29) are the same as those in equations (4.26) and (4.27), but the lower bounds are tighter in general. In fact, if $M^* \neq M$ (i.e., there are some channels with zero flow in the optimal solution), then the lower bounds in equations (4.26) and (4.27) are zero (certainly not very useful bounds).

We return to the study of formulation PF4. Let us find a lower bound on the average number of messages in the network at the optimal power point in a fashion similar to that for the M/M/1 series network of chapter 2. Since $\overline{N}_i = \rho_i/(1-\rho_i)$ for M/M/1, we have

$$\overline{N}_i = \rho_i + \frac{\rho_i^2}{1-\rho_i} = \rho_i + (\overline{N}_i)^2(1-\rho_i)$$

Therefore, summing over i yields

$$\sum_{i=1}^{M} \overline{N}_i = \sum_{i=1}^{M} \rho_i + \sum_{i=1}^{M} (\overline{N}_i)^2(1-\rho_i)$$

or

$$\sum_{i=1}^{M} \overline{N}_i = \sum_{i=1}^{M} \rho_i + \sum_{i=1}^{M} (\overline{N}_i)^2 - \sum_{i=1}^{M} \rho_i (\overline{N}_i)^2$$

The above equation holds for all values of γ. If we evaluate it at γ^* (the optimal power point) and then apply equation (4.21), we find that

$$\sum_{i=1}^{M} \rho_i^* = \sum_{i=1}^{M} \rho_i^* (\overline{N}_i^*)^2$$

This interesting equation relates the individual channel efficiencies and average number of messages at maximal power. As in chapter 2, we divide both sides of the above equation by $\rho_{\max}^*$ to yield

$$\sum_{i=1}^{M} \frac{\rho_i^*}{\rho_{\max}^*} = \sum_{i=1}^{M} \frac{\rho_i^*}{\rho_{\max}^*} (\overline{N}_i^*)^2$$

Since $\rho_i^*/\rho_{\max}^* \leq 1$ for $1 \leq i \leq M$, we have

$$\sum_{i=1}^{M} \frac{\rho_i^*}{\rho_{\max}^*} = (\overline{N}_i^*)^2$$

We use equation (4.21) to recognize the term on the right-hand side of the above inequality as the average number of messages in the network at maximum power. Thus we have the lower bound

$$\sum_{i=1}^{M} \frac{\rho_i^*}{\rho_{\max}^*} \leq \overline{N}^* \tag{4.30}$$

This is identical to equation (2.12) for the M/M/1 series network and equation (2.43) for the M/M/1 parallel network. We now write this lower bound in terms of known system parameters. Note that

$$\rho_i^* = \frac{\gamma^* R_i}{\mu C_i} \tag{4.31}$$

for all $1 \leq i \leq M$, and so

$$\sum_{i=1}^{M} \frac{\rho_i^*}{\rho_{\max}^*} = \sum_{i=1}^{M} \frac{R_i/C_i}{\max_{1\le j\le M} R_j/C_j} = \sum_{i=1}^{M} \frac{\min_{1\le j\le M} C_j/R_j}{C_i/R_i}$$

where we have used the fact that $R_i > 0$ for $1 \le i \le M$ for PF4 in order to obtain the last inequality. Thus we have the following

Theorem 4.4

For the power problem formulation PF4, the average number of messages in the system at maximal power satisfies

$$\sum_{i=1}^{M} \frac{\min_{1\le j\le M} C_j/R_j}{C_i/R_i} \le \overline{N}^* \tag{4.32}$$

Theorems 4.3 and 4.4 give upper and lower bounds on $\overline{N}^*$ for the formulation PF4 in terms of the given quantities of the particular network being considered. Note that Theorem 4.3 also holds for PF1, PF2 and PF3, while Theorem 4.4 holds for these three formulations if we replace M by M^* and R_i by R_i^*. However, this gives the lower bound in terms of unknown quantities which must be solved for (and then you would have the optimal answer anyway), and so it is not a useful bound. Also note that the lower bound in the form of equation (4.30) does hold for PF1, PF2 and PF3, but again it is given in terms of the unknowns ρ_i^*.

We now find bounds on the system throughput at optimal power for the problem PF4. Although the value of γ^* which yields the optimal power point may be determined by a root finding procedure, we are able to find bounds on this quantity (which is equal to the optimal scaling factor) in terms of the given channel capacities. (Of course, we see from equation (4.12) that $\gamma^* < \min_{1\le i\le M} \mu(C_i/R_i)$, because γ^* yields a feasible traffic pattern.) We use equation (4.31) to rewrite equation (4.27) as

$$\max_{1\le i\le M} \frac{\gamma^* R_i}{\mu C_i} \ge \frac{1}{2} \ge \min_{1\le i\le M} \frac{\gamma^* R_i}{\mu C_i}$$

This simplifies to

$$\frac{2}{\mu} \max_{1\le i\le M} \frac{1}{C_i/R_i} \ge \frac{1}{\gamma^*} \ge \frac{2}{\mu} \min_{1\le i\le M} \frac{1}{C_i/R_i}$$

or

$$\frac{2/\mu}{\min_{1\le i\le M} C_i/R_i} \ge \frac{1}{\gamma^*} \ge \frac{2/\mu}{\max_{1\le i\le M} C_i/R_i}$$

Inverting this equation reverses the signs of the inequalities, and we have the following

Theorem 4.5

For the power problem formulation PF4, the system throughput at maximal power (the optimal traffic level) satisfies

$$\frac{\mu}{2}\min_{1\le i\le M}\frac{C_i}{R_i} \le \gamma^* \le \frac{\mu}{2}\max_{1\le i\le M}\frac{C_i}{R_i} \tag{4.33}$$

This result may also be shown by the following argument. If we divide both sides of equation (4.19) by γ^*, we obtain

$$\sum_{i=1}^{M}\frac{1}{\mu(C_i/R_i)-\gamma^*} = \sum_{i=1}^{M}\frac{\gamma^*}{[\mu(C_i/R_i)-\gamma^*]^2}$$

or

$$\sum_{i=1}^{M}\frac{\mu(C_i/R_i)-\gamma^*}{[\mu(C_i/R_i)-\gamma^*]^2} = \sum_{i=1}^{M}\frac{\gamma^*}{[\mu(C_i/R_i)-\gamma^*]^2}$$

Collecting terms yields

$$\sum_{i=1}^{M}\frac{\mu(C_i/R_i)-2\gamma^*}{[\mu(C_i/R_i)-\gamma^*]^2} = 0 \tag{4.34}$$

We now use equation (4.34) to prove Theorem 4.5. First assume that

$$\gamma^* < \frac{\mu C_i}{2R_i} \qquad 1\le i\le M$$

In this case the sum on the left-hand side of equation (4.34) is positive, a contradiction. Therefore

$$\frac{\mu}{2}\min_{1\le i\le M}(C_i/R_i) \le \gamma^*$$

which is the left-hand inequality of Theorem 4.5. Similarly, if we assume that

$$\gamma^* > \frac{\mu C_i}{2R_i} \qquad 1\le i\le M$$

then the sum in equation (4.34) would be negative, another contradiction. Thus we also have

$$\gamma^* \le \frac{\mu}{2}\max_{1\le i\le M}(C_i/R_i)$$

which is the right-hand inequality of Theorem 4.5, completing this alternative proof.

Theorem 4.5 gives upper and lower bounds on γ^* in terms of the given channel capacities and ratios R_i. Note that if C_i/R_i are identical for all $1 \leq i \leq M$, then γ^* is equal to the common value $(\mu C_i)/(2R_i)$. In this case, equation (4.31) shows that $\rho_i^* = 1/2$ for all $1 \leq i \leq M$, and thus $\overline{N}_i^* = 1$. Recalling the discussion after Theorem 4.4, we observe that Theorem 4.5 may be applied to formulations PF1, PF2, and PF3; but the bounds will be in terms of M^* and the optimal ratios R_i^*, which are not known until the particular problem is solved.

4.1.3 Example Networks

We now consider several examples. We may apply the above theorems to the M/M/1 series network and to the M/M/1 parallel network with known routing which were analyzed in chapter 2, because both are instances of PF4. For the M/M/1 series network, we have $R_i = 1$ for all $1 \leq i \leq M$ (the total network traffic passes through each channel). In this case, the bounds given by Theorems 4.3 and 4.4 become (define $C_{\min} = \min_{1 \leq i \leq M} C_i$)

$$\sum_{i=1}^{M} \frac{C_{\min}}{C_i} \leq \overline{N}^* \leq M$$

while Theorem 4.5 reduces to (define $C_{\max} = \max_{1 \leq i \leq M} C_i$)

$$\frac{\mu}{2} C_{\min} \leq \gamma^* \leq \frac{\mu}{2} C_{\max}$$

These agree with the bounds given in chapter 2 for the M/M/1 series network.

For the M/M/1 parallel network with known routing (where the routing fractions correspond to p_i of chapter 2), we have $R_i = \lambda_i/\gamma = p_i$. Theorems 4.3 and 4.4 yield

$$\sum_{i=1}^{M} \frac{\min_{1 \leq j \leq M} C_j/p_j}{C_i/p_i} \leq \overline{N}^* \leq M$$

while Theorem 4.5 gives

$$\frac{\mu}{2} \min_{1 \leq i \leq M} \frac{C_i}{p_i} \leq \gamma^* \leq \frac{\mu}{2} \max_{1 \leq i \leq M} \frac{C_i}{p_i}$$

The bounds on $\overline{N}^*$ were also obtained in chapter 2 by using results proved for the M/G/1 parallel network. Here they were obtained by using results for the M/M/1 problem formulation PF4.

4.1.4 Analysis of PF3

We now use our results for the problem PF4 to study the formulation PF3. In this formulation, both the traffic level (for a given relative traffic matrix) and the routing are variables of optimization. Our strategy is to use a common technique of problem manipulation from mathematical programming which, in the terminology of [Geof70], is called *projection*. Simply stated, projection involves partitioning the original set of design variables (say into the sets X and Y), fixing y in Y and optimizing the resulting problem for x in X (which yields a function of y which we call $v(y)$), and then optimizing $v(y)$ over the set Y. In the problem PF3, the variables of optimization may be naturally partitioned into the routing variables $\{p_{ijk}\}$ and the traffic level α. Our solution strategy is to alternately solve for $\{p_{ijk}\}$ with α fixed, and then solve for α with $\{p_{ijk}\}$ fixed. This procedure illustrates possible interaction between routing and flow control, as we are simply alternating between finding the optimal routing (in terms of power) for a given flow and finding the optimal flow for a given routing. Such interplay between routing and flow control has been examined before, for example in [Gerl80a] in the context of a closed network model of a virtual circuit network.

Let us consider these two steps in our proposed solution procedure in more detail. The routing step involves finding the routing which maximizes power for a given traffic level α. In this case, the traffic matrix $\{\gamma_{jk}\}$ is constant, and the throughput of the network $\gamma = \alpha$ is also constant. Since $P = \gamma/T$ and γ is constant, maximizing power is equivalent to minimizing total average system delay T. Thus the routing step involves finding the routing which minimizes delay for a given traffic matrix. This problem is simply the flow assignment problem (FA) which has been discussed at length in [Frat73]. Any $\{p_{ijk}\}$ which minimizes T yields corresponding flows $\{\lambda_i\}$ from equation (4.2). These flows minimize T subject to the constraint that the λ_i represent a multicommodity flow. Thus we can recover optimal $\{\lambda_i\}$ from the optimal $\{p_{ijk}\}$. However, we cannot necessarily argue in the other direction; for if $\{\lambda_i\}$ minimizes T (subject to the constraint of representing a multicommodity flow), there may be several routings which yields such flow values. But using a solution procedure called the *flow deviation algorithm* described in [Frat73], such a routing can be obtained (in the form of a routing table) as a byproduct of the determination of the optimal flows. We wish to optimize with respect to the channel flows λ_i, because we obtain a convex cost multicommodity flow problem. As mentioned in chapter 1, this type of problem has been studied extensively, and several solution techniques are known. We use the flow deviation algorithm, since it may be easily modified to also yield routing tables (i.e., the $\{p_{ijk}\}$).

The flow control step is simply an instance of PF4, the solution of which we have discussed above. This problem is a single variable optimization problem, and it involves finding the root of equation (4.19) within the appropriate feasible interval. Thus it may be solved easily by any root finding procedure.

Let us then describe a strategy for attempting to solve problem PF3.

Initial Step

We first find a scaling factor $\alpha^{(0)}$ which yields an initial feasible solution. Unlike the case of some optimization problems (for example, the flow assignment problem), such an initial feasible solution always exists. We simply scale the given relative traffic matrix downward until a feasible traffic pattern is obtained. For example, set $\alpha = 1$ and arbitrarily assign all r_{jk} traffic to a fixed path from node j to node k. After this is done for all source-destination pairs (j, k), we scale the traffic down until all channel capacity constraints are met. (We have not considered the problem of the best way of choosing an initial feasible solution. An analysis of this type is valuable in terms of the convergence properties of an algorithm; we mention that such a study was conducted in [Cour80] for the flow assignment problem.) After $\alpha^{(0)}$ has been determined, calculate the power $P^{(0)}$. Iterations are now performed which consist of both a routing step and a flow control step. We proceed to the routing step of the first iteration.

Routing Step

The optimization problem to be solved is a convex cost multicommodity flow problem. For iteration $t+1$, perform the flow deviation algorithm to find the optimal routing for the given traffic matrix $\{\alpha^{(t)} r_{jk}\}$. Call the resulting routing fractions $\{p_{ijk}^{(t+1)}\}$. Calculate the ratios $R_i^{(t+1)} = \sum_j \sum_k p_{ijk}^{(t+1)} r_{jk}$, which are constant for this routing. Go to the flow control step for iteration $t+1$.

Flow Control Step

The optimization problem to be solved at this step is an instance of PF4. For iteration $t+1$, solve the single variable problem which involves finding the root of the polynomial given in equation (4.19). Call the resulting traffic level $\gamma^{(t+1)}$. Note that there is a unique root in the feasible interval $0 < \gamma < \min_{1 \le i \le M} \mu(C_i / R_i^{(t+1)})$, because P is a strictly concave function for problem PF4. Also note that $R_i^{(t+1)}$ are known quantities obtained from the previous routing step, and represent the fraction of external traffic routed over each channel i. We now check to see if the algorithm terminates.

Stopping Rule

Calculate the value of power for iteration $t+1$, which we call $P^{(t+1)}$. If $|P^{(t+1)} - P^{(t)}| < \epsilon$, where ϵ is a tolerance parameter, then stop. Otherwise, go to the routing step of the next iteration $t+2$.

Each iteration involves two steps. The routing step decreases total mean system delay under constant throughput and thus increases power. The flow control step either increases both throughput and delay or decreases both throughput and delay while increasing power (it searches for the knee of the (γ, T) curve defined in PF4). Thus after each iteration the value of power has increased. Figure 4.4 shows a family of throughput delay curves and a possible pattern of values for γ and T obtained from this algorithm.

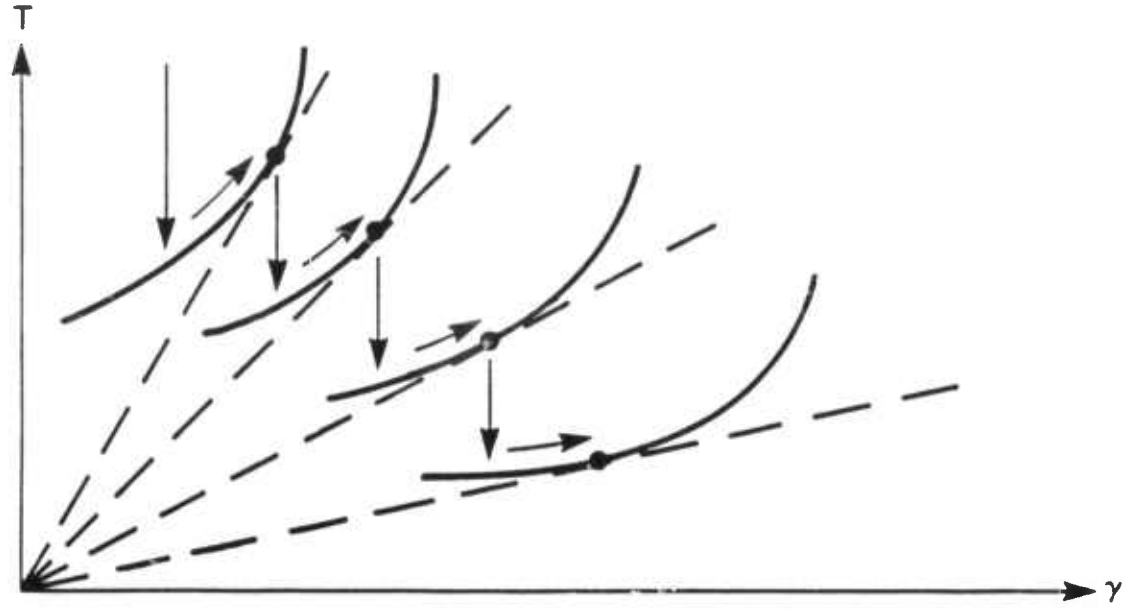

Figure 4.4 Convergence Path for PF3 Algorithm

The value of power at a point (γ, T) is the reciprocal of the slope of a line through the origin which passes through that point. Thus such lines are curves of constant power. Since the smaller the slope the higher the power, we wish to find points as near to the γ axis and as far from the T axis as possible which are on the knee of some throughput delay curve. Note from Figure 4.4 that for the routing step, we find the smallest T curve for that fixed throughput. Then for the flow control step, we simply find the knee of the T curve determined in the routing step. Thus we "go down" to the appropriate T curve and then slide along it (in one of the two directions) until we find its knee. We see that the resulting slopes of the lines through the origin to the knee of the curves decrease, and so the power increases. We now recall that the parallel net which modeled a single user with multiple paths is an instance of PF3, and so the results of chapter 3 apply. Thus the power function under consideration (for a particular instance of PF3) may not be concave or may have more than one critical point. Since the above algorithm searches for critical points, it may be necessary to perform it with several different feasible initial points.

Formulations PF1 and PF2 are much harder to solve. One approach is to reduce the problem to PF3 in the case of PF1 or to PF4 in the case of PF2. Various relative traffic matrices could be tried, and the corresponding values of power for each of them could be compared. Perhaps results for previous $\{r_{jk}\}$ could be used to help guide the choice of the next relative traffic matrix. In the case of formulation PF2, each relative traffic matrix yields an instance of PF4, which simply involves a single variable root finding problem. The polynomial given in equation (4.19) will always be used, but the fractions R_i will change with the choice of the relative traffic matrix. For PF1, each relative traffic matrix yields an instance of PF3. The algorithm described above involving both routing and flow control steps could thus be used in attempting to solve PF1. More needs to be done in analyzing these two more complicated optimization problems.

4.2 Other Definitions of Power

In the first part of this chapter, the problem of optimizing power was studied for networks with a general topology. Several optimization problems (denoted PF1, PF2, PF3, and PF4) were formulated with global network power $P = \gamma/T$ as the objective function in each case. We attacked the difficulty of the power problem for general network topologies by changing the problem formulation (constraints and/or decision variables) while keeping the same performance measure. In the remainder of this chapter we focus our attention on the objective function power itself. We review several different definitions of power which have appeared in the literature in addition to that of Giessler and extend them to general network settings. One particular performance measure, first introduced by Kleinrock, still has physical meaning as a throughput delay tradeoff function and still retains the desirable properties of power for the simple systems of chapter 2, but it yields a more satisfying solution for more general network problems.

4.2.1 A Continuum of Power Functions

In the previous sections of this chapter several optimization problems were studied for which power $P = \gamma/T$ was the objective function. Although formulations involving a relative traffic matrix were dealt with successfully, the most general problems (PF1 and PF2) remained difficult to solve. We now consider other power functions, all of which give the same results for the problem PF4 as those of P above, but which will hopefully allow us to attack the more difficult problems PF1 and PF2. Of course we also must preserve the desirable physical characteristics of P as a throughput-delay tradeoff function which, when maximized, directs us to operate at the knee of the (γ, T) curve.

We begin by reviewing several definitions of power which have previously appeared. The objective function which has been studied in chapters 2, 3, and 4 of this work for various network topologies and under different problem formulations was introduced by Giessler and his colleagues [Gies78]. It is simply defined as

$$P_G = \frac{\gamma}{T} \tag{4.35}$$

where γ is the system throughput and T is the mean system delay. Since throughput is a desirable quantity while delay is not, an appropriate performance measure for a computer network would be one which maximizes throughput and minimizes delay. We see that Giessler's definition of power incorporates both features into a single function.

Independently (and in fact prior to the work of Giessler) Nakamura and his colleagues introduced a power function in the paper [Yosh77] (published in Japanese) and analyzed it for several queueing systems. Their definition was motivated by a correspondence they developed between computer networks and electrical networks. The function which they introduced corresponded to power in an electrical network, and thus the new objective function was also called power. Nakamura and his colleagues have recently written a paper in English [Yosh81] which summarizes their results on power. We now motivate their power definition by considering a single server queueing system. The average delay T of a customer is the sum of W, the mean time spent waiting in queue (wasted time), plus $\bar{x}$, the average time spent actually being served (time spent in meaningful work). Thus a favorable objective is to maximize the fraction of useful time spent in the system $\bar{x}/T$, which is the same as minimizing the reciprocal $T/\bar{x}$. They therefore define power as

$$P_N = \frac{\gamma}{T/\bar{x}} = \frac{\gamma \bar{x}}{T} \tag{4.36}$$

so that (for a single server queueing system)

$$P_N = P_G \bar{x}$$

We note that neither of the above two power functions is normalized. A third definition of power which normalizes the parameters γ and T in a suitable fashion was introduced by Kleinrock in [Klei79] for a single server queueing system. In the context of a pure delay system, Kleinrock substituted the system efficiency ρ for the throughput γ and normalized the delay as $T/\bar{x}$. His definition is then

$$P_K = \frac{\rho}{T/\bar{x}} = \frac{\gamma(\bar{x})^2}{T} \tag{4.37}$$

so that (for a single server queueing system)

$$P_K = P_N \bar{x} = P_G(\bar{x})^2$$

Of course, for single server (pure delay) queueing systems, $\bar{x}$ is constant, and so points where the derivative with respect to γ is zero are the same for all three functions. Thus these three functions are maximized for the same value of γ at the knee of the (γ, T) curve. All the interesting relationships for single server systems derived above for P_G (such as Kleinrock's result that $\overline{N^*} = 1$ for M/G/1) also hold for P_N and P_K.

Let us try to extend these three definitions of power to arbitrary networks. We first note that Giessler's definition applies equally well to a general network. In fact Kleinrock [Klei78a] and Bharath-Kumar [Bhar80] studied P_G for tandems while Jaffe and Bharath-Kumar [Jaff81, Bhar81] analyzed it for more general networks. However, the other two power functions were defined only for a single node network (single server system). In order to extend P_N and P_K to an arbitrary network, we must determine the system parameter which naturally corresponds to the mean service time $\bar{x}$. We observe that, as a message travels through a computer network, the average time it spends in the network is the sum of the time it spends waiting in the various nodes for a free channel plus the time it spends actually being transmitted over the channels of the net. We may write $T = W + \bar{x}$ where W is the mean waiting time of a message and $\bar{x}$ is its mean service time. Thus $\bar{x}$ may be defined as the mean time it takes for a message to travel through the network if no other messages were present (i.e., if the network were empty). From this definition we have

$$\bar{x} = \sum_{i=1}^{M} \frac{\lambda_i}{\gamma} \bar{x}_i \tag{4.38}$$

where $\bar{x}_i$ is the mean time required to traverse the ith channel (so that $\bar{x}_i = \bar{b}/C_i = 1/\mu C_i$ for exponential message lengths of average $\bar{b} = 1/\mu$ bits). Of course we also have

$$W = \sum_{i=1}^{M} \frac{\lambda_i}{\gamma} W_i \tag{4.39}$$

where W_i is the mean waiting time at channel i. Adding equations (4.38) and (4.39) together, we obtain

$$T = W + \bar{x} = \sum_{i=1}^{M} \frac{\lambda_i}{\gamma} (W_i + \bar{x}_i)$$

which yields the familiar result [Klei64]

$$T = \sum_{i=1}^{M} \frac{\lambda_i}{\gamma} T_i \tag{4.40}$$

where T_i is the mean delay for channel i.

Using this definition of $\bar{x}$ we can extend P_N and P_K in the following way. We set

$$P_N = P_G \bar{x} = \frac{\gamma \bar{x}}{T}$$

which is, using equations (4.38) and (4.40),

$$P_N = \frac{\sum_{i=1}^{M} \lambda_i \bar{x}_i}{\sum_{i=1}^{M} \frac{\lambda_i}{\gamma} T_i}$$

By Little's result and the definition of ρ_i we have

$$P_N = \gamma \frac{\sum_{i=1}^{M} \rho_i}{\sum_{i=1}^{M} \bar{N}_i} \tag{4.41}$$

Similarly we write

$$P_K = P_G(\bar{\tau})^2 = P_N \bar{\tau}$$

Therefore, using equation (4.41), we have

$$P_K = \gamma \bar{\tau} \frac{\sum_{i=1}^{M} \rho_i}{\sum_{i=1}^{M} \bar{N}_i}$$

and so, by equation (4.38) and the definition of ρ_i,

$$P_K = \frac{\left(\sum_{i=1}^{M} \rho_i\right)^2}{\sum_{i=1}^{M} \bar{N}_i} \tag{4.42}$$

Thus, as in the single server case, we have

$$P_K = P_N \bar{\tau} = P_G(\bar{\tau})^2 \tag{4.43}$$

Therefore, any power problem solved previously with P_G as the objective function for which $\bar{\tau}$ (as defined in equation (4.38)) is *constant* also yields the same optimal power point for P_N and P_K (although the optimal function value will, in general, be different). In fact this argument also holds for the continuum of functions of the form

$$P_{(s)} = \frac{\gamma}{T}(\bar{\tau})^s \tag{4.44}$$

where s is a real number (note that $P_{(0)} = P_G$, $P_{(1)} = P_N$ and $P_{(2)} = P_K$). Since $\bar{\tau}$ is constant for a single node net (single server queueing system), all results derived above for P_G hold equally well for P_N and P_K (and any $P_{(s)}$). For example, all of these functions are maximized at the knee of the throughput-delay curve for single server systems, and Kleinrock's "keep the pipe full" result for the M/G/1 queueing system (see equation (2.4) of chapter 2) also holds when any of these other power functions are maximized.

We next observe that for a series network all messages traverse every channel in the tandem, and therefore $\bar{\tau}$ is constant for such a network topology as well ($\bar{\tau}$ is simply the sum of all the channel transmission times). Thus all our earlier results for series networks (both M/M/1 and M/D/1) also hold for P_N and P_K (and all $P_{(s)}$). Since $\bar{\tau}$ is constant for the

M/G/1 parallel network studied in chapter 2 ($\bar{x} = \sum_{i=1}^{M} p_i \bar{x}_i$), the results for this configuration also hold for these definitions of power. We now show that all results determined for Giessler's power function P_G for the formulation PF4 defined earlier in this chapter also hold for P_N and P_K. We need only show that $\bar{x}$, as defined in equation (4.38), is constant for problem PF4. Using the definition $R_i = \lambda_i/\gamma$ from equation (4.10) in equation (4.38) yields

$$\bar{x} = \sum_{i=1}^{M} R_i \bar{x}_i \tag{4.45}$$

Recalling that R_i is constant for PF4 we see that $\bar{x}$ is also constant for formulation PF4. Thus all results for PF4 (and those for PF1, PF2, and PF3 that had a proof depending on PF4) hold equally well with P_N or P_K (or $P_{(s)}$ for any real number s) as the objective function of the formulation. Therefore, we observe that Theorems 4.1, 4.2, 4.3, 4.4, and 4.5 are also true for these other performance measures.

We have seen that any of the above functions will be maximized at the knee of a (γ, T) curve. In the case of PF4, there is only one curve to consider, so that the maximum is identical when any of these objective functions is used. However, this is no longer the case for the other formulations PF1, PF2, and PF3. Although the global maximum will be at the knee of some throughput delay curve, the exponent on $\bar{x}$ in equation (4.44) will determine the globally optimal traffic and routing (i.e., which throughput delay curve is selected). Different choices of the objective function will give different curves; but always a knee of a (γ, T) curve will be chosen. Of course for topologies for which each user has only one possible routing (such as trees), problem PF3 reduces to PF4 and all performance measures will be maximized at the same point. We also note that by Theorems 4.1 and 4.3, for any formulation, P_N and P_K (and any $P_{(s)}$) are maximized at a point (which will be different for the different functions) where

$$\sum_{i=1}^{M} \bar{N}_i^* = \sum_{i=1}^{M} (\bar{N}_i^*)^2$$

and so

$$\bar{N}^* \leq M$$

for any of these power functions. But we observe that M^*, the number of channels with nonzero flow at the optimal solution, may be different for different definitions of power.

Along with the above properties that are preserved for all the various power functions, there are some properties which do not carry over. We first recall that the value of (Giessler's) power at a point (γ, T) on the throughput-delay curve is the reciprocal of the slope of a line through the origin to that point. We observe that this property is not retained by P_N or P_K. We see that the slope m of a line through the origin to a point on the (γ, T) curve satisfies

$$m = \frac{T}{\gamma} = \frac{\bar{x}}{P_N} = \frac{(\bar{x})^2}{P_K}$$

Although the value of P_K is no longer the reciprocal of the slope, we note that it still is maximized when the slope is minimized (i.e., at the knee of the curve). This is because $\bar{x}$ is constant and

$$P_K = \frac{1}{m}(\bar{x})^2$$

The same argument holds for P_N (or any $P_{(s)}$ with $s \neq 0$).

We next observe that the solution procedure for PF3 given earlier in this chapter cannot be applied to the functions P_N or P_K. This is because the analysis of the routing step which facilitated the use of the flow deviation algorithm is no longer applicable. That is, we no longer obtain a flow assignment problem as part of the routing step when these other definitions of power are used. For, consider the optimization problem of finding the routing $\{p_{i,jk}\}$ which maximizes (some notion of) power for a fixed traffic matrix $\{\gamma_{jk}\}$. For Giessler's definition it was observed that maximizing P_G in this case is equivalent to minimizing delay T, and so this problem reduces to the classical flow assignment problem with the flow deviation algorithm as a solution procedure. For problem PF3, the average service time $\bar{x}$ is no longer constant, and so the maximization of P_K (or P_N) under the assumption of a fixed traffic matrix is *not equivalent* to the minimization of T. Thus the problem (encountered in the routing step) of finding the optimal routing for a given traffic matrix is not equivalent to the flow assignment problem, and so the approach outlined above in this chapter for the solution of PF3 is no longer valid.

Let us illustrate this point with an example. Consider an M/M/1 parallel network with two channels, which represents a single user with two different paths for his messages. Let us study the problem of maximizing power for this simple network under the following assumptions. We assume that channel 1 has capacity $4C$ while channel 2 has capacity C. We also assume that the traffic matrix (which is equivalent to the throughput since there is only one user) is a given quantity and is equal to $\gamma_{12} = \gamma = (5\mu C)/2$. As discussed in section 4.1.1, if we use P_G as our performance measure, then the problem is simply an instance of the flow assignment problem. However, we now let P_K be our objective function, and consider two different routings of the traffic. If all of the traffic is put on the fast channel ($\lambda_1 = (5\mu C)/2$ and $\lambda_2 = 0$), then $T = 2/(3\mu C)$ and $\bar{x} = 1/(4\mu C)$. Therefore we find $P_K = 15/64$ for this routing. If we route the traffic so that the efficiency of both channels is 1/2 ($\lambda_1 = 2\mu C$ and $\lambda_2 = (\mu C)/2$), then $T = 4/(5\mu C)$ and $\bar{x} = 2/(5\mu C)$. In this case we find $P_K = 1/2$; although the average delay T is higher for the second routing, the value of P_K is also higher, unlike the behavior of P_G. We remark that, by results derived below, the latter routing where $\rho_i = 1/2$ for both channels is in fact globally optimal for P_K.

Despite some of the negative aspects given above when choosing a power function different from P_G, one particular choice yields desirable results for several formulations from the first part of this chapter. We choose P_K which generalizes Kleinrock's power function to an arbitrary network. We first rewrite equation (4.42) solely in terms of ρ_i. Since $\bar{N}_i = \rho_i/(1-\rho_i)$ for M/M/1, this equation becomes

$$P_K = \frac{\left(\sum_{i=1}^{M} \rho_i\right)^2}{\sum_{i=1}^{M} \frac{\rho_i}{1-\rho_i}} \tag{4.46}$$

Considered in terms of the channel utilizations ρ_i, the objective function P_K is identical for all formulations and all network topologies, unlike the power function P_G studied in chapters previously. This is because the throughput γ has "dropped out" of P_K in equation (4.46), while it is present in the expression for P_G. Thus the interdependencies of the ρ_i for a particular problem and network (due to flow conservation at the nodes, whether the relative traffic matrix and/or routing is given, etc.) appear in the function P_G but not in P_K. This is why the function P_G (considered as a function of ρ_i) will, in general, be different for different networks and formulations. If we use P_K as the power function, all networks and formulations give optimization problems with the same objective function; only the constraints of the optimization problem differ. We can thus compare various network topologies to each other when using P_K, whereas with P_G our interest is in comparing different possible operating points with respect to power for the same network.

The function P_K also occurs naturally from another point of view. Let us consider performance measures of the form

$$P_f = \frac{f[\{\gamma_{jk}\}]}{T/\bar{x}}$$

where f is a function of the traffic matrix. Note that if

$$f[\{\gamma_{jk}\}] = \sum_{j=1}^{N}\sum_{k=1}^{N} \gamma_{jk} = \gamma$$

then we obtain Nakamura's power function P_N. Let us try to choose an f from fairness considerations. We know that users whose messages traverse slow channels experience larger mean delay, and therefore may be assigned zero throughput in an optimal (unfair) solution. Toward obtaining a fair solution, let us weight the individual throughputs γ_{jk} by the time spent in transmission. We define $\bar{x}_{jk}$ to be the average time that the (j,k) traffic spends actually being transmitted over the various channels in its journey through the net. Thus

$$\bar{x}_{jk} = \sum_{i=1}^{M} p_{i,jk}\,\bar{x}_i$$

where, as above, $p_{i,jk}$ is the fraction of γ_{jk} traffic which uses channel i. Now define f by

$$f[\{\gamma_{jk}\}] \triangleq \sum_{j=1}^{N}\sum_{k=1}^{N}\bar{x}_{jk}\gamma_{jk}$$

Then we have

$$f[\{\gamma_{jk}\}] = \sum_{j=1}^{N}\sum_{k=1}^{N}\gamma_{jk}\sum_{i=1}^{M}p_{ijk}\bar{x}_i = \sum_{i=1}^{M}\bar{x}_i\sum_{j=1}^{N}\sum_{k=1}^{N}p_{ijk}\gamma_{jk}$$

which, using equations (4.2) and (4.38), becomes

$$f[\{\gamma_{jk}\}] = \sum_{i=1}^{M}\bar{x}_i\lambda_i = \gamma\bar{x}$$

Using this definition of f we have

$$P_f = \frac{f[\{\gamma_{jk}\}]}{T/\bar{x}} = \frac{\gamma\bar{x}}{T/\bar{x}} = \frac{\gamma}{T}(\bar{x})^2 = P_K$$

Thus fairness considerations have yielded Kleinrock's power function, and it seems a natural performance measure to examine.

4.2.2 Power of the M/M/1 Parallel Network (for P_K)

We now solve the parallel network problem of chapter 3 using the objective function P_K. We recall that this problem can be viewed as an instance of the most general formulation PF1 from the first part of this chapter. We also recall that the optimal solution of this network for P_G was unfair, was quite complicated to derive, and had the property that local power was not equal to global power. The model we consider is that of an M/M/1 parallel network with M channels and no restriction on the routing of messages through the net. We wish to find channel flows λ_i for $1 \le i \le M$ which maximize P_K. This yields a *multi-variable* optimization problem with M unknowns. We will find all critical points of P_K (points with $\nabla P_K = 0$) and compare them with the maximal boundary point as in chapter 3. To this end, we first find the M partial derivatives of P_K with respect to the channel flows. We may write equation (4.42) in terms of λ_i as

$$P_K = \frac{\left(\sum_{i=1}^{M}\lambda_i\bar{x}_i\right)^2}{\sum_{i=1}^{M}\lambda_i T_i}$$

Taking the partial derivative of P_K with respect to λ_j gives

$$\frac{\partial P_K}{\partial \lambda_j} = \frac{2\left(\sum_{i=1}^{M}\lambda_i T_i\right)\left(\sum_{i=1}^{M}\lambda_i\bar{x}_i\right)\bar{x}_j - \left(\sum_{i=1}^{M}\lambda_i\bar{x}_i\right)^2\left(\lambda_j\frac{dT_j}{d\lambda_j} + T_j\right)}{\left(\sum_{i=1}^{M}\lambda_i T_i\right)^2}$$

or

$$\frac{\partial P_K}{\partial \lambda_j} = \left[\frac{\sum_{i=1}^{M} \lambda_i \bar{x}_i}{\sum_{i=1}^{M} \lambda_i T_i}\right]^2 \left[\frac{2\sum_{i=1}^{M} \lambda_i T_i}{\sum_{i=1}^{M} \lambda_i \bar{x}_i} \bar{x}_j - \left(\lambda_j \frac{dT_j}{d\lambda_j} + T_j\right)\right]$$

If we assume that $\nabla P_K = 0$, then $\partial P_K / \partial \lambda_j = 0$ for all $1 \leq j \leq M$. Thus we have (using $\bar{x}_j = 1/\mu C_j$)

$$(\mu C_j)\left(\lambda_j \frac{dT_j}{d\lambda_j} + T_j\right) = \frac{2\sum_{i=1}^{M} \lambda_i T_i}{\sum_{i=1}^{M} \lambda_i \bar{x}_i} = (\mu C_k)\left(\lambda_k \frac{dT_k}{d\lambda_k} + T_k\right) \qquad \forall j, k \tag{4.47}$$

Since we assume that each channel acts as an M/M/1 queueing system, we may use equation (3.5) of chapter 3 which states that

$$\lambda_j \frac{dT_j}{d\lambda_j} + T_j = \frac{1}{\mu C_j} \cdot \frac{1}{(1-\rho_j)^2}$$

Substituting this expression into equation (4.47) yields

$$\frac{1}{(1-\rho_j)^2} = \frac{2\sum_{i=1}^{M} \lambda_i T_i}{\sum_{i=1}^{M} \lambda_i \bar{x}_i} = \frac{1}{(1-\rho_k)^2} \qquad \forall j, k \tag{4.48}$$

Therefore, for $\nabla P_K = 0$, we must have

$$\rho_j = \rho_k \qquad \forall j, k \tag{4.49}$$

and thus

$$\bar{N}_j = \bar{N}_k \qquad \forall j, k \tag{4.50}$$

Using equations (4.49) and (4.50), we find that equation (4.48) simplifies to

$$\frac{2M\,\bar{N}_1}{M\,\rho_1} = \frac{1}{(1-\rho_1)^2}$$

Since $\bar{N} = \rho/(1-\rho)$ for M/M/1, we find

$$\frac{2}{1-\rho_1} = \frac{1}{(1-\rho_1)^2}$$

or

$$\rho_1 = \frac{1}{2}$$

and so

$$\bar{N}_1 = 1$$

Thus $\rho_i = 1/2$ and $\bar{N}_i = 1$ for all $1 \leq i \leq M$ at this (unique) interior critical point of the function P_K, and the corresponding function value is

$$P_K = \frac{(M/2)^2}{M} = \frac{M}{4}$$

As in chapter 3, we must also find the optimal *boundary* point; this occurs when one of the λ_i is zero. Therefore we consider a parallel network with $M - 1$ channels, and by the above analysis, the optimal boundary point will be such that $\rho_i = 1/2$ for all these $M - 1$ channels. Thus the optimal boundary point will yield a function value of $P_K = (M - 1)/4$, and so the critical point of P_K is globally optimal.

We have shown the following

Theorem 4.6

For the M/M/1 parallel network (with unknown routing), the channel utilizations which maximize the power function P_K are

$$\rho_i^* = \frac{1}{2} \qquad \forall 1 \leq i \leq M \tag{4.51}$$

We also have the following

Theorem 4.7

For the M/M/1 parallel network (with unknown routing), the average number of messages at each channel when P_K is maximized is

$$\bar{N}_i^* = 1 \qquad \forall 1 \leq i \leq M \tag{4.52}$$

This theorem extends Kleinrock's "keep the pipe full" result to the M/M/1 parallel system with unknown routing. We note that the undesirable properties of the optimal solution for this network when P_G was maximized have here vanished. The optimal solution for P_K is fair (it is as fair as possible), while global power and local power are now the same. For P_K, the utilizations of the individual channels have taken on greater weight in the objective function. This is reflected in the optimal solution, because every channel is utilized at the same level regardless of the speed of the channel.

Let us now consider formulations PF1, PF2, and PF3 with objective function P_K for a *general* network topology. The optimization problem for the parallel network considered above is an instance of PF1, and the only constraints of this problem are that each ρ_i be between zero and one. We note that the feasible region for that particular problem consists of the set of all

M-dimensional vectors $\rho = (\rho_1, \ldots, \rho_M)$ such that $0 \leq \rho_i \leq 1$ for all $1 \leq i \leq M$, while the objective function is given solely in terms of the ρ_i in equation (4.46). Formulations PF1, PF2, and PF3 for general network topologies will have the same objective function as given in equation (4.46), but the feasible region will be a subset of the feasible region for the parallel network. The constraints imposed by the particular topology of the network and/or by the formulation being considered (given routing or relative traffic matrix) will reduce the set of ρ_i which are feasible. If we consider an arbitrary network and disregard these extra constraints, we obtain the parallel network optimization problem. If the optimal solution to the parallel network is feasible for the particular network we are examining, then it must be optimal for this other network also. That is, if the vector given by $\rho_i = 1/2$ for all i is *feasible* for a particular network problem, then it is *optimal*. This says that, if it is at all possible to "keep the pipe full" based upon the constraints imposed by the topology and/or formulation, then the optimal policy is to do just that. Also note that for such a network problem, $P_K^* = M/4$. Therefore, for any network with M channels and any formulation, we must have that $P_K^* \leq M/4$.

4.2.3 Examples

We now give several examples of network problems such that $\overline{N}_i = 1$ for all i is optimal ("keep the pipe full"). We have shown that the parallel network with arbitrary channel capacities is such an example, while the series network with equal channel capacities is another obvious example. A third interesting example is a unidirectional ring with M nodes and M channels, all channels having the same capacity C. We assume channel i connects nodes i and $i+1$ for $1 \leq i \leq M-1$, and that channel M connects node M and node 1 (the case of six nodes is illustrated in Figure 4.5).

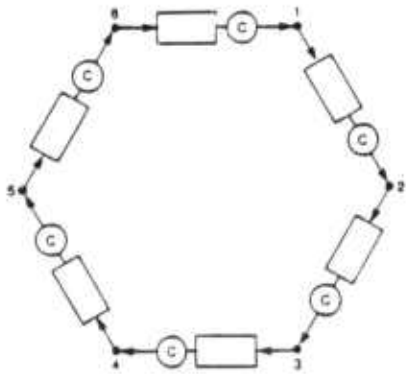

Figure 4.5 A Unidirectional Ring Network

Here we consider an instance of PF1; we wish to find the traffic matrix $\{\gamma_{jk}\}$ which optimizes P_K. Every source-destination pair (j, k) with $j \neq k$ corresponds to a user, so that there are $M(M-1)$ total users of the ring. Each user has a unique path that his messages take through the net. Thus no alternate routing is possible, and as discussed in chapter 4, PF1 reduces to PF2. We now ask if there is a traffic matrix which yields $\rho_i = 1/2$ for all channels i. If such a traffic matrix exists, it must be optimal by the above arguments.

Let us first consider a *uniform* traffic matrix (all γ_{jk} identical). We have that channel i connects node i to node $i+1$, and we now calculate the number of paths which use channel i. Since there is no traffic from a node to itself, there are $M-1$ paths using channel i for traffic originating at node i (one for every node other than node i). Similarly, there are $M-2$ paths which use channel i for traffic originating at node $i-1$. Continuing in this manner we see that there is one path for traffic from node $i+2$ (namely $i+2$ to $i+1$), and there is no path over channel i for traffic originating at node $i+1$. Thus the total number of paths using channel i is $1+\cdots+(M-1)=(M-1)M/2$. We want $\rho_i=1/2$, or equivalently $\lambda_i=(\mu C)/2$. Since λ_i is the sum of all traffic using channel i and $\{\gamma_{jk}\}$ is uniform, we must have

$$\lambda_i=\frac{(M-1)M}{2}\gamma_{jk}$$

where $1\le j\ne k\le M$. Equating the right-hand expression with the desired quantity $(\mu C)/2$ gives

$$\gamma_{jk}=\frac{\mu C}{(M-1)M}\qquad \forall j\ne k$$

Thus the external input to any node j is

$$\gamma_j=\sum_{k\ne j}\gamma_{jk}=\frac{\mu C}{M}\qquad \forall j$$

The total throughput is

$$\gamma=\sum_j\gamma_j=\sum_j\sum_{k\ne j}\gamma_{jk}=\mu C$$

Of course, by construction, we have

$$\lambda_i=\frac{\mu C}{2}\qquad \forall i$$

and

$$\rho_i=\frac{1}{2}\qquad \forall i$$

Thus we have found a (uniform) traffic matrix $\{\gamma_{jk}\}$ which yields a "keep the pipe full" solution, and so by the above discussion it must globally maximize P_K for the formulation PF1. Note that it is also the global optimum for problem PF3 (which is equivalent to PF4 since there is no alternate routing) when the given relative traffic matrix for the ring is uniform. Note also that the optimal solution is fair, since all users receive (the same) non-zero throughput. The optimal traffic pattern is shown in Figure 4.6.

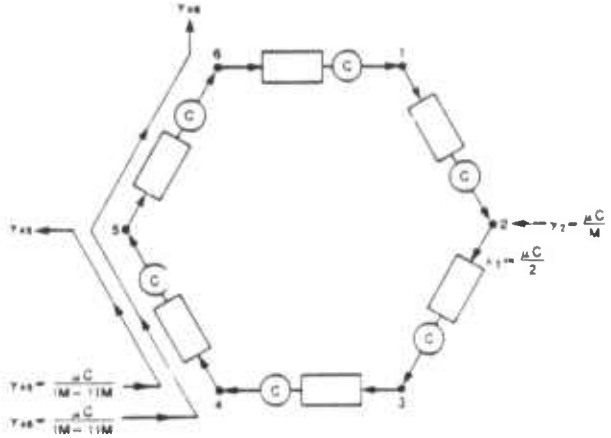

Figure 4.6 A Fair Optimal Traffic Pattern

Although the above (uniform) traffic matrix optimizes P_K for PF1, it is not the only optimal solution. A second optimal point is given by the non-uniform traffic matrix

$$\gamma_{jk} = \begin{cases} \dfrac{\mu C}{2} & k = j(\text{mod } M) + 1 \\ 0 & \text{otherwise} \end{cases}$$

In this case, we clearly have

$$\lambda_i = \frac{\mu C}{2} \qquad \forall i$$

and

$$\rho_i = \frac{1}{2} \qquad \forall i$$

Thus this traffic matrix is also optimal for PF1 with P_K as the objective function. Note that only those M users who are sending messages to adjacent nodes have non-zero throughput. Therefore, although this optimal solution does "keep the pipe full", it is also *unfair*. The optimal traffic pattern for this "selfish" solution is shown in Figure 4.7.

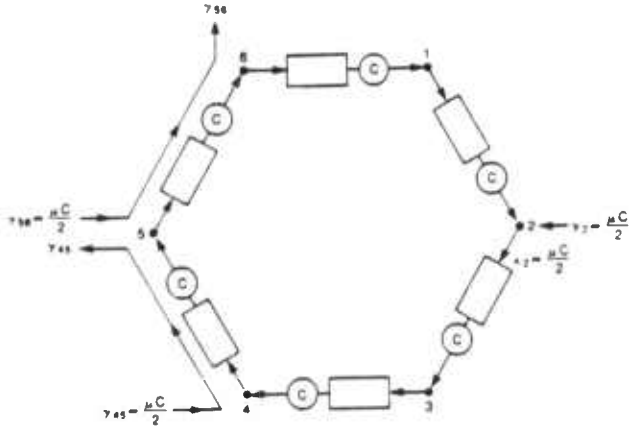

Figure 4.7 An Unfair Optimal Traffic Pattern

We have given several examples of networks which have $\rho_i = 1/2$ for all i as the optimal solution using objective function P_K. No doubt the reader can think of other such networks for which "keep the pipe full" is optimal. If we assume that we have M channels, each channel having the same capacity C, then the channels may be arranged in a parallel system, a tandem, or a ring and the resulting networks will have the same optimal solution when P_K is maximized. In the following table we compare the values of several system parameters at this optimal point for the three different networks.

	parallel	series	ring (uniform)	ring (selfish)
γ	$\frac{M}{2}(\mu C)$	$\frac{1}{2}(\mu C)$	μC	$\frac{M}{2}(\mu C)$
T	$\frac{2}{\mu C}$	$\frac{2M}{\mu C}$	$\frac{M}{\mu C}$	$\frac{2}{\mu C}$
$\bar{x}$	$\frac{1}{\mu C}$	$\frac{M}{\mu C}$	$\frac{M/2}{\mu C}$	$\frac{1}{\mu C}$
P_G	$\frac{M}{4}(\mu C)^2$	$\frac{1}{4M}(\mu C)^2$	$\frac{1}{M}(\mu C)^2$	$\frac{M}{4}(\mu C)^2$
P_N	$\frac{M}{4}(\mu C)$	$\frac{1}{4}(\mu C)$	$\frac{1}{2}(\mu C)$	$\frac{M}{4}(\mu C)$
P_K	$\frac{M}{4}$	$\frac{M}{4}$	$\frac{M}{4}$	$\frac{M}{4}$

Table 4.1

This table shows the effect of the term $\bar{x}$ in the objective function. Note that all three systems have the same fraction $(1/2)$ of useful work to total delay $\bar{x}/T$. This is no surprise since

$$\frac{\bar{x}}{T} = \frac{\gamma \bar{x}}{\gamma T} = \frac{\sum_{i=1}^{M} \rho_i}{\sum_{i=1}^{M} N_i} = \frac{M/2}{M} = \frac{1}{2}$$

In this section we have examined several definitions of power. We have seen that choosing a measure first introduced by Kleinrock yields the rule of thumb that (to maximize P_K) one should always operate a network in order to "keep the pipe full" as long as the constraints of the network topology and problem formulation allow this to be achieved. However, a solution which maximizes P_K and does "keep the pipe full" may still be unfair.

CHAPTER 5
Generalized Power for Simple Networks

In the previous three chapters of this dissertation we have studied several types of power functions for various network topologies and several optimization problem formulations. Each of these power performance measures incorporated a throughput-delay tradeoff in the form γ/T. In this chapter we examine an extension of the power definition which was first introduced by Kleinrock. This new tradeoff function (actually a family of functions) allows the analyst to favor throughput over delay or vice versa. We study this family in terms of formulations which yield simple optimization problems. In the next chapter we extend our analysis to multi-variable optimization problems of the type studied in chapter 3.

5.1 Generalised Power

In chapters 2, 3, and 4 of this work we have examined the throughput-delay tradeoff functions P_G, P_N and P_K. Each of these three power functions is an increasing function of throughput and a decreasing function of delay. However, the relative importance of throughput γ and mean delay T is given by the ratio γ/T in all three cases. Noting this restriction on the relationship of throughput to delay imposed in these definitions of power, Kleinrock [Klei79] introduced a family of functions indexed by the continuous parameter r which enables the analyst to vary the importance of throughput relative to delay depending on the particular application. This new tradeoff function, called *generalized power,* was defined simply as

$$P_{K,r} = \frac{\rho^r}{T/\bar{x}} \tag{5.1}$$

where r is a positive real number. The restriction that r be positive insures that this new performance measure is an increasing function of throughput. For, suppose that $r \leq 0$, and define $r' \triangleq -r$ so that $r' \geq 0$. Then $P_{K,r} = 1/\gamma^{r'}T$, and maximizing $P_{K,r}$ would be equivalent to *minimizing* $\gamma^{r'}T$; the generalized power function would then no longer be an increasing function of the throughput γ, and the optimal solution would be to let $\gamma = 0$, clearly an unacceptable situation. Thus we will always restrict r to be positive.

Although the definition of $P_{K,r}$ is an extension of the power function P_K, we may also define a generalized power family based on P_G as

$$P_{G,r} = \frac{\gamma^r}{T} \tag{5.2}$$

Note that large values of r indicate that throughput is preferred, while smaller values of r are

used when the minimization of delay is of prime importance. Also note that

$$P_{K,r} = \frac{(\gamma\bar{x})^r}{T/\bar{x}} = \frac{\gamma^r}{T}(\bar{x})^{r+1} = P_{G,r}(\bar{x})^{r+1}$$

so that both extensions yield the same optimal solutions of problem formulations for which $\bar{x}$ is constant. In this chapter, we concentrate on such optimization problems, and thus it is immaterial which generalized power definition we use. We choose to examine the family $P_{G,r}$ and, for notational convenience, abbreviate it as P_r in the remainder of this chapter. In chapter 6 we will distinguish the particular generalized power definition used for a particular problem. We first review several results on generalized power which have appeared in the literature. We then analyze various networks in terms of the optimization problems of chapter 4 with this new objective function. Many of the known results about generalized power are obtained as special cases of the theorems which we prove below.

5.2 Previous Work on Generalized Power

We begin with a short summary of some selected results about generalized power. This family of functions was analyzed by Kleinrock for several queueing systems, and his analysis was extended to more general computer network models by Jaffe and Bharath-Kumar. Let us briefly state some of the interesting results for generalized power from these papers and indicate the source of each result. All of the results which we now review are applicable to either of the above generalized power functions, since the network problems studied have constant $\bar{x}$.

Kleinrock [Klei79] showed that

$$\bar{N}^* = r \tag{5.3}$$

and

$$\rho^* = \frac{r}{r+1} \tag{5.4}$$

when generalized power P_r is maximized for the queueing system M/M/1. Note that equation (5.3) extends Kleinrock's M/M/1 result that $\bar{N}^* = 1$ for ordinary power ($r = 1$). However, the corresponding analogue of the "keep the pipe full" result for M/G/1 does not hold for arbitrary r; Kleinrock [Klei79] gave the complicated formula for $\bar{N}^*$ for M/G/1. He also showed

$$\lim_{r\to\infty} \frac{\bar{N}^*}{r} = \frac{1+\nu^2}{2} \tag{5.5}$$

for M/G/1, where ν is the coefficient of variation of the service time distribution (standard deviation divided by the mean). Kleinrock also found that, for large r,

$$\rho^* \cong 1 - \frac{1}{r} \tag{5.6}$$

for the M/G/1 queueing system.

5.2.1 The M/M/1 Series Network

Bharath-Kumar [Bhar80] extended these results on generalized power to the M/M/1 series network. He showed that, for arbitrary channel capacities, the equation

$$r\sum_{i=1}^{M}\overline{N}_i^* = \sum_{i=1}^{M}(\overline{N}_i^*)^2 \tag{5.7}$$

holds at the optimal generalized power point. From this equation he obtained the upper bound

$$\overline{N}^* \leq Mr \tag{5.8}$$

for the average number in the series system when P_r is maximized.

For an M/M/1 series net with *equal* capacities, Bharath-Kumar was able to show that

$$\overline{N}_i^* = r \qquad 1 \leq i \leq M \tag{5.9}$$

and that

$$\rho_i^* = \frac{r}{r+1} \qquad 1 \leq i \leq M \tag{5.10}$$

Of course this yields

$$\overline{N}^* = Mr \tag{5.11}$$

for the case of equal channel capacities.

5.3 Optimality Conditions

We will extend these results for generalized power by Kleinrock and Bharath-Kumar to more complicated network topologies and problem formulations. Before we do this, it behooves us to obtain conditions which insure that a point is optimal with respect to generalized power for a "simple" network problem formulation. These conditions were first given by Kleinrock, and may be derived in a manner similar to that of chapter 2 for ordinary power.

Consider a single server queueing system (a single node net). As in previous chapters, we consider the feasible region of possible values of throughput γ (which is identical to the input rate for this system) for our objective function P_r to be $0 \leq \gamma \leq 1/\overline{x}$ (i.e., $0 \leq \rho \leq 1$). Since P_r is positive on the interior of the feasible region and is zero at the two endpoints, we see that its maximum must be attained at an interior point. Assuming P_r is differentiable, by familiar arguments the derivative $dP_r/d\gamma$ must be zero at the optimal point. By differentiating equation (5.2) we see that the optimal value of throughput γ^* satisfies Kleinrock's equation

$$rT^* = \gamma^* \left.\frac{dT}{d\gamma}\right|_{\gamma=\gamma^*} \tag{5.12}$$

Note that the optimal point no longer occurs at the knee of the throughput delay curve (unless $r=1$), and that the reciprocal of the slope of a line through the origin to a point on the throughput delay curve does not give the value of generalized power at the point (again unless $r=1$). By Little's result, equation (5.12) is equivalent to

$$r\overline{N}^* = (\gamma^*)^2 \frac{dT}{d\gamma}\Big|_{\gamma=\gamma^*} \tag{5.13}$$

We observe that equations (5.12) and (5.13) hold whether or not P_r has "nice" properties (such as concavity). Although we show below that generalized power is generally not concave, this fact does not matter in the above characterization of the optimal generalized power point. We note that equations (5.12) and (5.13) characterize the optimal generalized power point, not only for a single server queueing system, but also for any of the simple single variable optimization problems studied previously (such as PF4 of chapter 4).

5.4 Non-Concavity of Generalized Power

It was erroneously stated in [Bhar81] that generalized power is concave for the M/M/1 series network. We now show that generalized power is not concave (for $r > 1$) even for the M/M/1 queueing system (a series net with one channel). Since $T = 1/(\mu C - \gamma)$ for M/M/1 (where the mean service time is $\bar{x} = 1/\mu C$), we have

$$P_r = \frac{\gamma^r}{T} = \gamma^r(\mu C - \gamma)$$

or

$$P_r = \gamma^r \mu C - \gamma^{r+1}$$

Differentiating the above equation with respect to γ yields

$$\frac{dP_r}{d\gamma} = r\gamma^{r-1}\mu C - (r+1)\gamma^r$$

Note that by setting $dP_r/d\gamma = 0$, we obtain

$$\gamma^* = \frac{r}{r+1}(\mu C)$$

This yields equation (5.4) and thus also equation (5.3). Differentiating again with respect to γ gives

$$\frac{d^2P_r}{d\gamma^2} = r(r-1)\gamma^{r-2}\mu C - (r+1)r\gamma^{r-1}$$

or

$$\frac{d^2P_r}{d\gamma^2} = r\gamma^{r-2}[(r-1)\mu C - (r+1)\gamma] \tag{5.14}$$

For $r < 1$, we observe that $d^2P_r/d\gamma^2 < 0$ for $0 < \gamma < \mu C$, and so P_r is (strictly) concave for all feasible γ ($\rho < 1$). For $r = 1$, the second derivative is identically equal to -2, and thus again P_r is (strictly) concave for all γ. For $r > 1$, the behavior of P_r is more complicated. From equation (5.14), we see that $d^2P_r/d\gamma^2 > 0$ for $\gamma < [(r-1)\mu C]/(r+1)$. Thus P_r is (strictly) *convex* for γ in this region, which shows that generalized power is not a concave function for M/M/1 if $r > 1$. We also note that P_r is (strictly) concave for $[(r-1)\mu C]/(r+1) < \gamma$.

Graphs of P_r versus γ for $r = 1/4$, $r = 1$, and $r = 4$ for the system M/M/1 are shown below in Figure 5.1, where we have chosen $\mu C = 1$ for convenience.

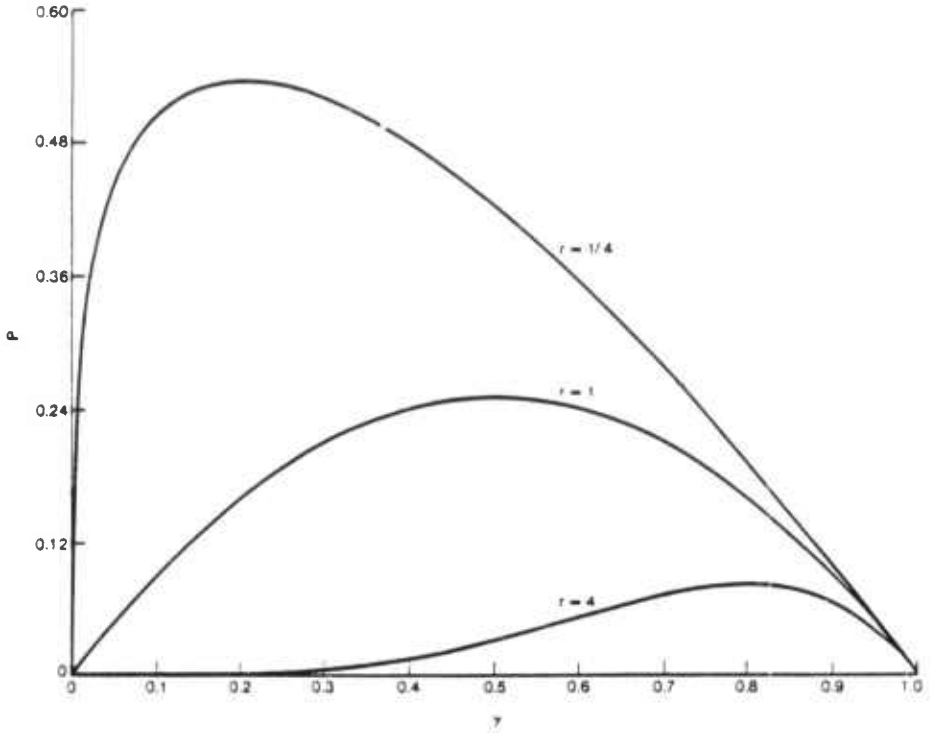

Figure 5.1 Generalized Power for M/M/1

We see that $P_{1/4}$ is strictly concave with $\gamma^* = 1/5$, while P_1 is also strictly concave with $\gamma^* = 1/2$. However, P_4 is strictly convex for $0 \leq \gamma \leq 3/5$ and strictly concave for $3/5 \leq \gamma \leq 1$, with $\gamma^* = 4/5$ (where its derivative with respect to γ is zero). We do note that P_4 is unimodal. These graphs illustrate that, although generalized power is not necessarily concave, equation (5.12) (or equation (5.13)) characterizes the optimal solution for simple single variable optimization problems with the maximization of P_r as the objective. This is true even if P_r is not unimodal; although there may be several points which satisfy equation (5.12), the global optimum must be one of them. We now apply these equations to the solution of various such problems.

5.5 The M/D/1 Series Network

We first consider a series net of M channels with arbitrary channel capacities and constant message length which was analyzed in chapter 2 for ordinary power. Recall that this model may be used to represent a path that a message takes through a computer network, since the length of the message does not change as it travels through the net. We will use equation (5.13) to determine the throughput (equivalent to the input rate) γ^* that maximizes the objective function P_r. We proceed in a manner analogous to the derivation of equation (2.23). As noted in chapter 2, from a result of Rubin [Rubi74], the average delay in the M/D/1 tandem is given by

$$T = \frac{\gamma(\bar{x}_{max})^2}{2(1-\gamma\bar{x}_{max})} + \sum_{i=1}^{M}\bar{x}_i$$

where $\bar{x}_{max} = \bar{b}/C_{min}$ is the maximum of the M mean channel transmission times (see equation (2.19)). Here C_{min} is the capacity of the slowest channel of the tandem. Therefore, we have

$$\frac{dT}{d\gamma} = \left[\frac{\bar{x}_{max}}{1-\gamma\bar{x}_{max}}\right]^2 \cdot \frac{1}{2}$$

as in equation (2.20). Substituting this into equation (5.13) yields

$$r\bar{N}^* = \left[\frac{\rho^*_{max}}{1-\rho^*_{max}}\right]^2 \cdot \frac{1}{2} \tag{5.15}$$

where $\rho_{max} = \gamma\bar{x}_{max} = (\gamma\bar{b})/C_{min}$ is the maximum of the M channel utilizations. From the expression for the mean waiting time W for this system, equation (5.15) is also equivalent to

$$r\bar{N}^* = \frac{\gamma^* W^*}{1-\rho^*_{max}}$$

or

$$r\bar{N}^* = \frac{\bar{N}^* - \sum_{i=1}^{M}\rho_i^*}{1-\rho^*_{max}} \tag{5.16}$$

Equation (5.16) yields

$$r\bar{N}^* - r\bar{N}^*\rho^*_{max} = \bar{N}^* - \sum_{i=1}^{M}\frac{\rho_i^*}{\rho^*_{max}}\cdot\rho^*_{max}$$

Since $\rho_i^* = \gamma^*\bar{b}/C_i$ for all $1 \le i \le M$, and therefore $\rho_i^*/\rho^*_{max} = C_{min}/C_i$, we have

$$(r-1)\bar{N}^* = \left(r\bar{N}^* - \sum_{i=1}^{M}\frac{C_{min}}{C_i}\right)\rho^*_{max}$$

For the case of ordinary power ($r = 1$), we obtain Theorem 2.4 of chapter 2.

We now assume that $r \neq 1$. Therefore

$$\rho^*_{max} = \frac{(r-1)\overline{N}^*}{r\overline{N}^* - \sum_{i=1}^{M} \frac{1}{\alpha_i}} \tag{5.17}$$

where we have written $C_i = \alpha_i C_{min}$ as in chapter 2 (so that $\alpha_i \geq 1$). We also have

$$1 - \rho^*_{max} = \frac{\overline{N}^* - \sum_{i=1}^{M} \frac{1}{\alpha_i}}{r\overline{N}^* - \sum_{i=1}^{M} \frac{1}{\alpha_i}} \tag{5.18}$$

Since $0 < \rho^*_{max} < 1$, equations (5.17) and (5.18) yield

$$\overline{N}^* > \sum_{i=1}^{M} \frac{1}{\alpha_i} \quad \text{for} \quad r > 1 \tag{5.19}$$

and

$$\overline{N}^* < \sum_{i=1}^{M} \frac{1}{\alpha_i} \quad \text{for} \quad r < 1 \tag{5.20}$$

Thus

$$\lim_{r \to \infty} r\overline{N}^* = \infty \tag{5.21}$$

and

$$\lim_{r \to 0} r\overline{N}^* = 0 \tag{5.22}$$

We obtain an exact expression for $\overline{N}^*$ (depending on r) as follows. Using equations (5.17) and (5.18), we find that equation (5.15) becomes

$$r\overline{N}^* = \left[\frac{(r-1)\overline{N}^*}{\overline{N}^* - \sum_{i=1}^{M} \frac{1}{\alpha_i}} \right]^2 \cdot \frac{1}{2}$$

which yields

$$\left(\overline{N}^* - \sum_{i=1}^{M} \frac{1}{\alpha_i}\right)^2 = \frac{(r-1)^2}{r} \overline{N}^* \cdot \frac{1}{2} \tag{5.23}$$

This quadratic equation may be solved to give an exact expression for $\overline{N}^*$ in terms of r and the known system parameters α_i. We write equation (5.23) in the form

$$(\overline{N}^*)^2 - \left[2\sum_{i=1}^{M} \frac{1}{\alpha_i} + \frac{(r-1)^2}{2r}\right] \overline{N}^* + \left(\sum_{i=1}^{M} \frac{1}{\alpha_i}\right)^2 = 0$$

which yields the two roots

$$\overline{N}^* = \sum_{i=1}^{M} \frac{1}{\alpha_i} + \frac{(r-1)^2}{4r} \pm \left[\frac{(r-1)^2}{2r} \sum_{i=1}^{M} \frac{1}{\alpha_i} + \left[\frac{(r-1)^2}{4r} \right]^2 \right]^{\frac{1}{2}}$$

Therefore, we have

$$\overline{N}^* - \sum_{i=1}^{M} \frac{1}{\alpha_i} = \frac{(r-1)^2}{4r} \pm \left[\frac{(r-1)^2}{2r} \sum_{i=1}^{M} \frac{1}{\alpha_i} + \left[\frac{(r-1)^2}{4r} \right]^2 \right]^{\frac{1}{2}} \tag{5.24}$$

Now note that, for $r > 0$ ($r \neq 1$), the right-hand side of equation (5.24) is strictly positive if the positive square root is chosen, and it is strictly negative if the negative square root is chosen. Also note that, from equations (5.19) and (5.20), the left-hand side of equation (5.24) is strictly positive for $r > 1$ and is strictly negative for $r < 1$. From these two observations we conclude that the positive square root must be chosen when $r > 1$, and the negative square root must be chosen when $r < 1$. Thus, for $r > 1$,

$$\overline{N}^* = \sum_{i=1}^{M} \frac{1}{\alpha_i} + \frac{(r-1)^2}{4r} + \left[\frac{(r-1)^2}{2r} \sum_{i=1}^{M} \frac{1}{\alpha_i} + \left[\frac{(r-1)^2}{4r} \right]^2 \right]^{\frac{1}{2}} \tag{5.25}$$

while for $r < 1$

$$\overline{N}^* = \sum_{i=1}^{M} \frac{1}{\alpha_i} + \frac{(r-1)^2}{4r} - \left[\frac{(r-1)^2}{2r} \sum_{i=1}^{M} \frac{1}{\alpha_i} + \left[\frac{(r-1)^2}{4r} \right]^2 \right]^{\frac{1}{2}} \tag{5.26}$$

We now determine the limiting behavior of $\overline{N}^*$ as $r \to \infty$. We first rewrite equation (5.23) as

$$\left(\overline{N}^* - \sum_{i=1}^{M} \frac{1}{\alpha_i} \right)^2 = \left(1 - \frac{1}{r}\right)^2 r\overline{N}^* \cdot \frac{1}{2} \tag{5.27}$$

By equation (5.21), the right-hand side of equation (5.27) becomes infinite as $r \to \infty$, and thus the left-hand side does also. This shows that

$$\lim_{r \to \infty} \overline{N}^* = \infty \tag{5.28}$$

We can describe this behavior in greater detail as follows. Expanding the left-hand side of equation (5.27) and then dividing both sides of the result by $r\overline{N}^*$ yields

$$\frac{\overline{N}^*}{r} - \frac{2\sum_{i=1}^{M} \frac{1}{\alpha_i}}{r} + \frac{\left(\sum_{i=1}^{M} \frac{1}{\alpha_i} \right)^2}{r\overline{N}^*} = \left(1 - \frac{1}{r}\right)^2 \cdot \frac{1}{2}$$

Using equation (5.21), we have the following

Theorem 5.1

For the M/D/1 series network, the average number of messages in system when generalized power P_r is maximized satisfies

$$\lim_{r\to\infty}\frac{\overline{N}^*}{r}=\frac{1}{2} \tag{5.29}$$

If there is only one channel ($M=1$), we have an M/D/1 queueing system. In this case, Theorem 5.1 agrees with Kleinrock's result for M/G/1 given by equation (5.5), since the coefficient of variation satisfies $\nu=0$ for constant service time.

However, equation (5.29) holds for any (fixed) number of channels M. Since the mean delay T is invariant with respect to the order of the channels in the tandem, let us assume that the first channel is the slowest. For this channel configuration, queueing occurs at the first channel only. As r increases, throughput becomes favored over delay in P_r. Thus $\overline{N}^*$ grows large as more and more messages are sent through the tandem in order to increase the throughput. Since all waiting takes place at the first channel, we have $\overline{N}_i^*=\rho_i^*$ for $2\le i\le M$, and so these quantities are bounded by 1 no matter how large r becomes. Thus almost all of the contribution to $\overline{N}^*$ comes from the first channel, and the series network looks like a single M/D/1 system (in terms of $\overline{N}^*$) as $r\to\infty$. This explains the agreement of Theorem 5.1 with Kleinrock's M/G/1 limiting result even for $M>1$.

We next determine the limiting behavior of $\overline{N}^*$ as $r\to 0$. We first rewrite equation (5.23) as

$$r\Big(\overline{N}^*-\sum_{i=1}^{M}\frac{1}{\alpha_i}\Big)=\frac{\overline{N}^*}{\overline{N}^*-\sum_{i=1}^{M}\frac{1}{\alpha_i}}\cdot\frac{(r-1)^2}{2} \tag{5.30}$$

By equation (5.22), the left-hand side of equation (5.30) approaches zero as $r\to 0$, and thus the right-hand side does also. This shows that

$$\lim_{r\to 0}\overline{N}^*=0 \tag{5.31}$$

We can describe this behavior in greater detail as follows. We first write equation (5.23) as

$$\frac{\overline{N}^*}{r}=\frac{2}{(r-1)^2}\Big(\overline{N}^*-\sum_{i=1}^{M}\frac{1}{\alpha_i}\Big)^2$$

Using equation (5.31), we have the following

Theorem 5.2

For the M/D/1 series network, the average number of messages in system when generalized power P_r is maximized satisfies

$$\lim_{r \to 0} \frac{\overline{N}^*}{r} = 2\left(\sum_{i=1}^{M} \frac{1}{\alpha_i}\right)^2 \tag{5.32}$$

If there is only one channel ($M = 1$), we have an M/D/1 queueing system, and Theorem 5.2 becomes

$$\lim_{r \to 0} \frac{\overline{N}^*}{r} = 2 \tag{5.33}$$

A second variable of particular interest is $\rho^*_{\max}$. Once $\overline{N}^*$ is determined from equation (5.23), the value of $\rho^*_{\max}$ is given by equation (5.17). Another expression for $\rho^*_{\max}$ may be found by using equation (5.15). Thus we have

$$2r\overline{N}^*(1 - \rho^*_{\max})^2 = (\rho^*_{\max})^2$$

and so, by taking square roots,

$$\rho^*_{\max} = \frac{\sqrt{2r\overline{N}^*}}{1 + \sqrt{2r\overline{N}^*}} \tag{5.34}$$

To find the limiting behavior of $\rho^*_{\max}$ as $r \to \infty$, we note that, by equation (5.18),

$$r(1 - \rho^*_{\max}) = \frac{r\overline{N}^* - r\sum_{i=1}^{M} \frac{1}{\alpha_i}}{r\overline{N}^* - \sum_{i=1}^{M} \frac{1}{\alpha_i}}$$

or

$$r(1 - \rho^*_{\max}) = \frac{1 - \frac{1}{\overline{N}^*}\sum_{i=1}^{M} \frac{1}{\alpha_i}}{1 - \frac{1}{r\overline{N}^*}\sum_{i=1}^{M} \frac{1}{\alpha_i}}$$

Using equations (5.21) and (5.28), we obtain the limiting result

$$\lim_{r \to \infty} r(1 - \rho^*_{\max}) = 1 \tag{5.35}$$

Therefore, for large values of r, we have

$$\rho^*_{\max} \cong 1 - \frac{1}{r} \tag{5.36}$$

This agrees with Kleinrock's equation (5.6) for M/G/1 in the case of one channel ($M = 1$).

To find the limiting behavior of ρ^*_{max} as $r \to 0$, we note that, by equation (5.17),

$$\frac{\rho^*_{max}}{r} = \frac{r-1}{r\overline{N}^* - \sum_{i=1}^{M} \frac{1}{\alpha_i}} \left(\frac{\overline{N}^*}{r}\right)$$

Using equations (5.22) and (5.32), we obtain the limiting result

$$\lim_{r \to 0} \frac{\rho^*_{max}}{r} = 2\sum_{i=1}^{M} \frac{1}{\alpha_i} \tag{5.37}$$

Therefore, for small values of r, we have

$$\rho^*_{max} \cong 2r\sum_{i=1}^{M} \frac{1}{\alpha_i} \tag{5.38}$$

In the case of one channel ($M = 1$), we have an M/D/1 queueing system, and equation (5.38) reduces to

$$\rho^* \cong 2r \tag{5.39}$$

The values of other system parameters at optimal generalized power may also be found once $\overline{N}^*$ and ρ^*_{max} have been evaluated. Recall that $\rho_i^* = (\gamma^* \overline{b})/C_i$ and so $\rho^*_{max} = (\gamma^* \overline{b})/C_{min}$. Therefore, using equation (5.34),

$$\rho_i^* = \frac{\rho_i^*}{\rho^*_{max}} \cdot \rho^*_{max} = \frac{C_{min}}{C_i}\left[\frac{\sqrt{2r\overline{N}^*}}{1+\sqrt{2r\overline{N}^*}}\right]$$

for $1 \leq i \leq M$. The optimal value of throughput is

$$\gamma^* = \frac{\rho^*_{max}}{\overline{b}/C_{min}} = \frac{C_{min}}{\overline{b}}\left[\frac{\sqrt{2r\overline{N}^*}}{1+\sqrt{2r\overline{N}^*}}\right]$$

and the optimal value of mean system delay is

$$T^* = \frac{\overline{N}^*}{\gamma^*} = \frac{\overline{b}/C_{min}}{\rho^*_{max}}\overline{N}^* = \frac{\overline{b}}{C_{min}}\left[\frac{\sqrt{2r\overline{N}^*} + 2r\overline{N}^*}{2r}\right]$$

The optimal generalized power itself is

$$P_r^* = P_r(\gamma^*) = \frac{(\gamma^*)^r}{T^*} = \frac{(\gamma^*)^{r+1}}{\overline{N}^*} = \left[\frac{C_{min}}{\overline{b}} \cdot \frac{\sqrt{2r\overline{N}^*}}{1+\sqrt{2r\overline{N}^*}}\right]^{r+1} \cdot \frac{1}{\overline{N}^*}$$

We leave it to the reader to determine the value of other variables of interest at optimal generalized power (such as W^*, W_i^*, T_i^*, and $\overline{N}_i^*$).

We remark that the pleasing result $\overline{N}^* = M$ when ordinary power is maximized for the equal channel capacity case of the M/D/1 series network does not extend to generalized power. Note that such an extension in the form $\overline{N}^* = Mr$ was shown to hold for the M/M/1 series network by Bharath-Kumar, and it is listed above as equation (5.11). We also note that the limiting results for $\overline{N}^*/r$ and ρ^*_{max} for the M/D/1 series network are invariant under scaling of channel capacities and the ordering of the channels. However, although the limiting results for $r \to \infty$ in equations (5.28) and (5.34) are the same regardless of the number of channels M, the limiting results for $r \to 0$ in equations (5.31) and (5.36) do depend on how many channels there are.

5.6 The M/G/1 Parallel Network (Known Routing)

We next turn our attention to the M/G/1 parallel network with known routing studied in chapter 2 for ordinary power. We recall that quantities p_i are given which represent the fraction of traffic which is transmitted on the ith channel and which satisfy $0 < p_i < 1$ and $\sum_{i=1}^{M} p_i = 1$. We may also regard p_i as the probability that a random message entering the parallel net will use channel i. The flow of messages on the ith channel is Poisson and has rate $\lambda_i = p_i\gamma$ where γ is the total system throughput.

The analysis of generalized power for this system proceeds in a manner similar to that for the M/D/1 series network. In chapter 2 it was shown that the average delay of a message for this parallel net is given by

$$T = \sum_{i=1}^{M} p_i W_i + \sum_{i=1}^{M} p_i \overline{x}_i = \sum_{i=1}^{M} p_i \frac{p_i \gamma (\overline{x}_i)^2}{1 - p_i \gamma \overline{x}_i} \left(\frac{1 + \nu_i^2}{2}\right)$$

where ν_i is the coefficient of variation of the service time distribution for the ith channel. Therefore, we have

$$\frac{dT}{d\gamma} = \sum_{i=1}^{M} p_i \frac{p_i (\overline{x}_i)^2}{(1 - p_i \gamma \overline{x}_i)^2} \left(\frac{1 + \nu_i^2}{2}\right)$$

Using $\rho_i = \lambda_i \overline{x}_i = p_i \gamma \overline{x}_i$ and equation (5.13), we have

$$r\overline{N}^* = \sum_{i=1}^{M} \left[\frac{\rho_i^*}{1 - \rho_i^*}\right]^2 \left(\frac{1 + \nu_i^2}{2}\right) \tag{5.40}$$

We now recall (see equation (2.34) of chapter 2) that

$$\overline{N}_i = \rho_i + \frac{(\rho_i)^2}{1 - \rho_i} \left(\frac{1 + \nu_i^2}{2}\right)$$

for M/G/1. Substituting this expression into equation (5.40) yields

$$r\sum_{i=1}^{M}\rho_i^* + r\sum_{i=1}^{M}\frac{(\rho_i^*)^2}{1-\rho_i^*}\left(\frac{1+\nu_i^2}{2}\right) = \sum_{i=1}^{M}\left[\frac{\rho_i^*}{1-\rho_i^*}\right]^2\left(\frac{1+\nu_i^2}{2}\right)$$

or

$$r\sum_{i=1}^{M}\rho_i^* = \sum_{i=1}^{M}\left[\frac{\rho_i^*}{1-\rho_i^*}\right]^2\left(\frac{1+\nu_i^2}{2}\right)[1-r(1-\rho_i^*)] \tag{5.41}$$

Using $\rho_i^* = \lambda_i^* \bar{x}_i = \gamma^* p_i \bar{x}_i$ for $1 \le i \le M$, we see that equation (5.41) is equivalent to a polynomial equation involving the known quantities p_i, $\bar{x}_i$ and ν_i (and r) and the single unknown γ^*. Thus γ^* may be found by any polynomial root finding procedure as was the case for ordinary power (see chapter 2).

Equation (5.40) is also equivalent to

$$r\bar{N}^* = \sum_{i=1}^{M}\frac{\lambda_i^* W_i^*}{1-\rho_i^*}$$

or

$$r\bar{N}^* = \sum_{i=1}^{M}\frac{\bar{N}_i^* - \rho_i^*}{1-\rho_i^*} \tag{5.42}$$

We will use equations (5.40) and (5.42) to obtain results concerning the optimal solution of P_r under various assumptions about the parallel system we are studying.

5.6.1 The M/M/1 Parallel Network (Known Routing)

We first assume that the distribution of message length is exponential with mean $\bar{b} = 1/\mu$. Thus each channel acts as an M/M/1 queueing system, and so we have an M/M/1 parallel network for which the routing (p_i) is given. Since $\bar{N}_i = \rho_i/(1-\rho_i)$ and $\nu_i = 1$ for all $1 \le i \le M$, equation (5.40) becomes

$$r\sum_{i=1}^{M}\bar{N}_i^* = \sum_{i=1}^{M}(\bar{N}_i^*)^2 \tag{5.43}$$

which is the same as equation (5.7) of Bharath-Kumar for the M/M/1 series network! We will see later in Theorem 5.5 that equations (5.7) and (5.43) are instances of a more general result.

We now use this equation to obtain an upper bound on the average number of messages in the network at optimal generalized power. Bharath-Kumar employed such an argument, but we give a simplified proof similar to that of equation (4.24). From equation (5.43) we have

$$r\bar{N}^* = r\sum_{i=1}^{M}\bar{N}_i^* = 2r\sum_{i=1}^{M}\bar{N}_i^* - \sum_{i=1}^{M}(\bar{N}_i^*)^2$$

Adding and subtracting Mr^2 from the right-hand side yields

$$r\bar{N}^* = Mr^2 - \sum_{i=1}^{M}\left[r^2 - 2r\bar{N}_i^* + (\bar{N}_i^*)^2\right]$$

or

$$r\bar{N}^* = Mr^2 - \sum_{i=1}^{M}(r - \bar{N}_i^*)^2 \tag{5.44}$$

This gives us the bound

$$\bar{N}^* \leq Mr \tag{5.45}$$

for the M/M/1 parallel network (with known routing). This is identical to equation (5.8) of Bharath-Kumar for the M/M/1 series network. We will see later in Theorem 5.7 that equations (5.8) and (5.45) are immediate consequences of that more general result.

We may also rewrite equation (5.44) in the form

$$r(Mr - \bar{N}^*) = \sum_{i=1}^{M}(r - \bar{N}_i^*)^2$$

so that

$$r\sum_{i=1}^{M}(r - \bar{N}_i^*) = \sum_{i=1}^{M}(r - \bar{N}_i^*)^2 \tag{5.46}$$

This yields a "dual" to equation (5.43), which is also a particular case of Theorem 5.6 below.

For an M/M/1 series network with equal channel capacities, Bharath-Kumar was able to show that equations (5.9), (5.10), and (5.11) hold. The analogous situation for the M/M/1 parallel network is the case of equal loads on all M channels. This will depend not only on the channel capacities, but also on the fraction of traffic which uses the ith channel. Since

$$\rho_i = \lambda_i \bar{x}_i = \frac{p_i \gamma}{\mu C_i}$$

the case of equal loads occurs when the (given) ratios p_i/C_i are equal for $1 \leq i \leq M$. If the channels have equal utilizations, the average number of messages $\bar{N}_i$ at the individual channels will be the same for all i for the case of M/M/1. Equation (5.43) thus becomes

$$r\,M\bar{N}_i^* = M(\bar{N}_i^*)^2$$

for any $1 \leq i \leq M$. Therefore, if p_i/C_i are equal for all i for the M/M/1 parallel network (which is the equal load case), then

$$\bar{N}_i^* = r \qquad 1 \leq i \leq M \tag{5.47}$$

and

$$\rho_i^* = \frac{r}{r+1} \qquad 1 \le i \le M \tag{5.48}$$

Of course this yields

$$\overline{N}^* = Mr \tag{5.49}$$

5.6.2 Equal Loads (Arbitrary Service Time Distributions)

The above equations (5.47), (5.48), and (5.49) characterize the optimal solution when P_r is maximized for the M/M/1 parallel network with equal loads. We now study the M/G/1 parallel network under the same assumption of equal channel utilizations. We set $\rho_i \triangleq \rho$ for $1 \le i \le M$. In this case, equation (5.42) may be rewritten as

$$r\overline{N}^*(1-\rho^*) = \overline{N}^* - M\rho^*$$

or

$$(r-1)\overline{N}^* = (r\overline{N}^* - M)\rho^*$$

For the case of ordinary power ($r = 1$) we obtain Theorem 2.8 of chapter 2.

We now assume that $r \neq 1$. Therefore

$$\rho^* = \frac{(r-1)\overline{N}^*}{r\overline{N}^* - M} \tag{5.50}$$

and so

$$1 - \rho^* = \frac{\overline{N}^* - M}{r\overline{N}^* - M} \tag{5.51}$$

Since $0 < \rho^* < 1$, equations (5.50) and (5.51) yield

$$\overline{N}^* > M \quad \text{for} \quad r > 1 \tag{5.52}$$

and

$$\overline{N}^* < M \quad \text{for} \quad r < 1 \tag{5.53}$$

Thus

$$\lim_{r \to \infty} r\overline{N}^* = \infty \tag{5.54}$$

and

$$\lim_{r \to 0} r\overline{N}^* = 0 \tag{5.55}$$

We obtain an exact expression for $\overline{N}^*$ (depending on r) as follows. Since we are considering the case of equal loads, equation (5.40) is

$$r\overline{N}^* = \left[\frac{\rho^*}{1-\rho^*}\right]^2 \sum_{i=1}^{M}\left(\frac{1+\nu_i^2}{2}\right) \tag{5.56}$$

Using equations (5.50) and (5.51), we find that equation (5.54) becomes

$$r\overline{N}^* = \left[\frac{(r-1)\overline{N}^*}{\overline{N}^* - M}\right]^2 \sum_{i=1}^{M}\left(\frac{1+\nu_i^2}{2}\right)$$

which yields

$$(\overline{N}^* - M)^2 = \frac{(r-1)^2}{r}\,\overline{N}^* \sum_{i=1}^{M}\left(\frac{1+\nu_i^2}{2}\right) \tag{5.57}$$

This quadratic equation may be solved to give an exact expression for $\overline{N}^*$ in terms of the number of channels M, the coefficients of variation ν_i for the service time distribution at the individual channels, and the parameter r. Note that equation (5.57) is similar to equation (5.23) for the M/D/1 series network, and we may proceed as in the derivation of equations (5.25) and (5.26). We write equation (5.57) in the form

$$(\overline{N}^*)^2 - \left[2M + \frac{(r-1)^2}{r}\sum_{i=1}^{M}\left(\frac{1+\nu_i^2}{2}\right)\right]\overline{N}^* + M^2 = 0$$

which yields the two roots

$$\overline{N}^* = M + \frac{(r-1)^2}{2r}\sum_{i=1}^{M}\left(\frac{1+\nu_i^2}{2}\right) \pm \left[M\,\frac{(r-1)^2}{r}\sum_{i=1}^{M}\left(\frac{1+\nu_i^2}{2}\right) + \left[\frac{(r-1)^2}{2r}\sum_{i=1}^{M}\left(\frac{1+\nu_i^2}{2}\right)\right]^2\right]^{\frac{1}{2}}$$

Therefore, we have

$$\overline{N}^* - M = \frac{(r-1)^2}{2r}\sum_{i=1}^{M}\left(\frac{1+\nu_i^2}{2}\right) \pm \left[M\,\frac{(r-1)^2}{r}\sum_{i=1}^{M}\left(\frac{1+\nu_i^2}{2}\right) + \left[\frac{(r-1)^2}{2r}\sum_{i=1}^{M}\left(\frac{1+\nu_i^2}{2}\right)\right]^2\right]^{\frac{1}{2}} \tag{5.58}$$

Now note that, for $r > 0$ ($r \neq 1$), the right-hand side of equation (5.58) is strictly positive if the positive square root is chosen, and it is strictly negative if the negative square root is chosen. Also note that, from equations (5.52) and (5.53), the left-hand side of equation (5.58) is strictly positive for $r > 1$ and is strictly negative for $r < 1$. From these two observations we conclude that the positive square root must be chosen when $r > 1$, and the negative square root must be chosen when $r < 1$. Thus, for $r > 1$,

$$\overline{N}^* = M + \frac{(r-1)^2}{2r}\sum_{i=1}^{M}\left(\frac{1+\nu_i^2}{2}\right) + \left[M\,\frac{(r-1)^2}{r}\sum_{i=1}^{M}\left(\frac{1+\nu_i^2}{2}\right) + \left[\frac{(r-1)^2}{2r}\sum_{i=1}^{M}\left(\frac{1+\nu_i^2}{2}\right)\right]^2\right]^{\frac{1}{2}} \tag{5.59}$$

while for $r < 1$

$$\overline{N}^* = M + \frac{(r-1)^2}{2r}\sum_{i=1}^{M}\left(\frac{1+\nu_i^2}{2}\right) - \left[M\frac{(r-1)^2}{r}\sum_{i=1}^{M}\left(\frac{1+\nu_i^2}{2}\right) + \left[\frac{(r-1)^2}{2r}\sum_{i=1}^{M}\left(\frac{1+\nu_i^2}{2}\right)\right]^2\right]^{\frac{1}{2}} \quad (5.60)$$

We now determine the limiting behavior of $\overline{N}^*$ as $r \to \infty$ in a manner similar to that for the M/D/1 series network. We first rewrite equation (5.57) as

$$(\overline{N}^* - M)^2 = \left(1 - \frac{1}{r}\right)^2 r\overline{N}^* \sum_{i=1}^{M}\left(\frac{1+\nu_i^2}{2}\right) \quad (5.61)$$

Using equation (5.54), we note that

$$\lim_{r\to\infty} \overline{N}^* = \infty \quad (5.62)$$

Dividing both sides of equation (5.61) by $r\overline{N}^*$ yields

$$\frac{\overline{N}^*}{r} - \frac{2M}{r} + \frac{M^2}{r\overline{N}^*} = \left(1 - \frac{1}{r}\right)^2 \sum_{i=1}^{M}\left(\frac{1+\nu_i^2}{2}\right)$$

Using equation (5.54), we have the following

Theorem 5.3

For the M/G/1 parallel network with equal loads, the average number of messages in the system when generalized power P_r is maximized satisfies

$$\lim_{r\to\infty} \frac{\overline{N}^*}{r} = \sum_{i=1}^{M}\left(\frac{1+\nu_i^2}{2}\right) \quad (5.63)$$

The case $M = 1$ (one channel) is Kleinrock's equation (5.5) for the queueing system M/G/1.

We next determine the limiting behavior of $\overline{N}^*$ as $r \to 0$. We first rewrite equation (5.57) as

$$r(\overline{N}^* - M) = \frac{\overline{N}^*}{\overline{N}^* - M} \cdot (r-1)^2 \sum_{i=1}^{M}\left(\frac{1+\nu_i^2}{2}\right) \quad (5.64)$$

By equation (5.55), the right-hand side of equation (5.64) must approach zero as $r \to 0$, and so

$$\lim_{r\to 0} \overline{N}^* = 0 \quad (5.65)$$

We write equation (5.57) as

$$\frac{\overline{N}^*}{r} = \frac{(\overline{N}^* - M)^2}{(r-1)^2 \sum_{i=1}^{M}\left(\frac{1+\nu_i^2}{2}\right)}$$

Using equation (5.65), we have the following

Theorem 5.4

For the M/G/1 parallel network with equal loads, the average number of messages in the system when generalized power P_r is maximized satisfies

$$\lim_{r\to 0} \frac{\overline{N}^*}{r} = \frac{M^2}{\sum_{i=1}^{M}\left(\frac{1+\nu_i^2}{2}\right)} \tag{5.66}$$

If there is only one channel ($M = 1$), we have an M/G/1 queueing system, and Theorem 5.4 becomes

$$\lim_{r\to 0} \frac{\overline{N}^*}{r} = \frac{2}{1+\nu^2} \tag{5.67}$$

We now examine the behavior of the common optimal channel utilization ρ^* for this equal load situation. Equation (5.56) may be rewritten as

$$r\overline{N}^*(1-\rho^*)^2 = (\rho^*)^2 \sum_{i=1}^{M}\left(\frac{1+\nu_i^2}{2}\right)$$

Taking square roots yields

$$\rho^* = \frac{\sqrt{r\overline{N}^*}}{\sqrt{r\overline{N}^*} + \sqrt{\sum_{i=1}^{M}(1+\nu_i^2)/2}} \tag{5.68}$$

To find the limiting behavior of ρ^* as $r \to \infty$, we note that, by equation (5.51),

$$r(1-\rho^*) = \frac{r\overline{N}^* - rM}{r\overline{N}^* - M}$$

or

$$r(1-\rho^*) = \frac{1 - \frac{M}{\overline{N}^*}}{1 - \frac{M}{r\overline{N}^*}}$$

Using equations (5.54) and (5.62), we obtain the limiting result

$$\lim_{r\to\infty} r(1-\rho^*) = 1 \tag{5.69}$$

which is similar to that for the M/D/1 series network. Thus for large values of r, we have

$$\rho^* \cong 1 - \frac{1}{r} \tag{5.70}$$

The case $M = 1$ (one channel) is Kleinrock's equation (5.6) for the queueing system M/G/1.

To find the limiting behavior of ρ^* as $r \to 0$, we note that, by equation (5.50),

$$\frac{\rho^*}{r} = \frac{r-1}{r\bar{N}^* - M}\left(\frac{\bar{N}^*}{r}\right)$$

Using equations (5.55) and (5.66), we obtain the limiting result

$$\lim_{r \to 0} \frac{\rho^*}{r} = \frac{M}{\sum_{i=1}^{M}\left(\frac{1+\nu_i^2}{2}\right)} \tag{5.71}$$

Therefore, for small values of r, we have

$$\rho^* \cong \frac{Mr}{\sum_{i=1}^{M}\left(\frac{1+\nu_i^2}{2}\right)} \tag{5.72}$$

In the case of one channel ($M = 1$), we have an M/G/1 queueing system, and equation (5.72) reduces to

$$\rho^* \cong \frac{2r}{1+\nu^2} \tag{5.73}$$

The values of other system parameters at optimal generalized power (such as throughput, delay, and generalized power itself) may be found once the values ρ^* and $\bar{N}^*$ are determined by proceeding as in the analysis of the M/D/1 series network.

Let us examine the case of identical coefficients of variation for service time at all M channels. We set $\nu_i \triangleq \nu$ for $1 \leq i \leq M$. Then the limiting result of Theorem 5.3 becomes

$$\lim_{r \to \infty} \frac{\bar{N}^*}{r} = M\left(\frac{1+\nu^2}{2}\right) \tag{5.74}$$

while the limiting result of Theorem 5.4 becomes

$$\lim_{r \to 0} \frac{\bar{N}^*}{r} = \frac{2M}{1+\nu^2} \tag{5.75}$$

Since we have equal loads and equal coefficients of variation, the average number of messages $\bar{N}_i$ at the ith channel will be the same for all M channels for the case of M/G/1. We thus have $\bar{N}_i^* = \bar{N}^*/M$ for $1 \leq i \leq M$, and so equation (5.74) yields

$$\lim_{r \to \infty} \frac{\bar{N}_i^*}{r} = \frac{1+\nu^2}{2} \tag{5.76}$$

while equation (5.75) yields

$$\lim_{r \to 0} \frac{\bar{N}_i^*}{r} = \frac{2}{1+\nu^2} \tag{5.77}$$

These results agree with equations (5.5) and (5.67), as they must.

5.7 Arbitrary M/M/1 Network

In this section we first extend our analysis of generalized power to the single variable optimization problem formulation PF4 which was introduced in chapter 4. We study an M/M/1 network with arbitrary topology, but we assume that the external traffic supplied by each user is known as a fraction of the total input, and that the paths that a user's messages take as well as the fraction of his traffic on each path are also known. In the notation of the previous chapter, the relative traffic matrix $\{r_{jk}\}$ and the routing $\{p_{ijk}\}$ are given. For this situation, the ratios $R_i = \lambda_i/\gamma$ were shown to be constant. We will use equation (5.13) to determine the optimal traffic level γ^* which maximizes generalized power P_r.

As shown in equation (4.17), the derivative of the average system delay T with respect to γ satisfies

$$\frac{dT}{d\gamma} = \sum_{i=1}^{M} R_i^2 T_i^2$$

for an M/M/1 network under formulation PF4. Multiplying this equation by γ^2, and then using equation (5.13) and Little's result, we have

$$r\overline{N}^* = r\sum_{i=1}^{M} \overline{N}_i^* = \sum_{i=1}^{M} (\overline{N}_i^*)^2 \tag{5.78}$$

Arguing as in chapter 4, we use (for M/M/1)

$$\overline{N}_i = \frac{R_i\gamma}{\mu C_i - R_i\gamma} = \frac{\gamma}{\mu(C_i/R_i) - \gamma}$$

to rewrite equation (5.78) in terms of the unknown variable γ as

$$r\sum_{i=1}^{M} \frac{\gamma^*}{\mu(C_i/R_i) - \gamma^*} = \sum_{i=1}^{M} \left[\frac{\gamma^*}{\mu(C_i/R_i) - \gamma^*}\right]^2 \tag{5.79}$$

Thus the optimal value of throughput is simply the root of a polynomial in γ (equivalent to equation (5.79) above) which lies within the open interval $(0, \min_{1\le i\le M} \mu[C_i/R_i])$. The root may be determined by any standard root finding procedure.

Equation (5.78) holds for any instance of PF4; thus equation (5.7) for the M/M/1 series network and equation (5.43) for the M/M/1 parallel network are both special cases. We may extend this result to the M/M/1 problem formulations PF1, PF2 and PF3 of chapter 4 by arguing as in the proof of Theorem 4.1. We thus have the following

Theorem 5.5

For the four generalized power problem formulations (PF1, PF2, PF3, and PF4), the average number in the M/M/1 network at maximum generalized power satisfies

$$r\bar{N}^* = r\sum_{i=1}^{M} \bar{N}_i^* = \sum_{i=1}^{M} (\bar{N}_i^*)^2 \tag{5.80}$$

We now use equation (5.80) to show that the dual equation (5.46) holds for these four formulations. We first rewrite equation (5.80) in the form

$$\sum_{i=1}^{M} \bar{N}_i^* (r - \bar{N}_i^*) = 0 \tag{5.81}$$

Defining $Q_i \triangleq r - \bar{N}_i^*$, we have

$$\sum_{i=1}^{M} (r - Q_i) Q_i = 0$$

or

$$r\sum_{i=1}^{M} Q_i = \sum_{i=1}^{M} (Q_i)^2$$

This immediately yields the following

Theorem 5.6

For the four generalized power problem formulations (PF1, PF2, PF3, and PF4), the average number in the M/M/1 network at maximum generalized power satisfies

$$r\sum_{i=1}^{M} (1 - \bar{N}_i^*) = \sum_{i=1}^{M} (1 - \bar{N}_i^*)^2 \tag{5.82}$$

We next use equation (5.82) to yield the bound of equations (5.8) and (5.45) for these four problem formulations. Rewriting equation (5.82) as

$$r(Mr - \bar{N}^*) = \sum_{i=1}^{M} (r - \bar{N}_i^*)^2$$

and noting that the right-hand side of the above equation is non-negative, we obtain

$$r(Mr - \bar{N}^*) \geq 0$$

Dividing by the positive quantity r, we have shown the following

Theorem 5.7

For the four generalized power problem formulations (PF1, PF2, PF3, and PF4), the average number in an M/M/1 network (with M channels) at maximum generalized power satisfies

$$\overline{N}^* \leq Mr \tag{5.83}$$

This generalizes the upper bound given in Theorem 4.3 for ordinary power ($r = 1$). However, the lower bound obtained in Theorem 4.4 for P has thus far resisted a similar generalization for P_r.

Following the approach of chapter 4 for ordinary power, equation (5.81) may be used to obtain bounds on the individual $\overline{N}_i^*$. Since at least one $\overline{N}_i^*$ must be positive, if $\overline{N}_i^* < r$ for all i, then the sum in equation (5.81) would be positive. Similarly, if $\overline{N}_i^* > r$ for all i, then the sum in equation (5.81) would be negative. Thus we have shown that

$$\max_{1 \leq i \leq M} \overline{N}_i^* \geq r \geq \min_{1 \leq i \leq M} \overline{N}_i^* \tag{5.84}$$

Since $\overline{N}_i = \rho_i/(1-\rho_i)$ for M/M/1, we see that equation (5.84) gives

$$\max_{1 \leq i \leq M} \rho_i^* \geq \frac{r}{r+1} \geq \min_{1 \leq i \leq M} \rho_i^* \tag{5.85}$$

Recall that these bounds can be strengthened for PF1, PF2, and PF3 by substituting $\overline{M}^*$, the number of channel with nonzero flow at optimal generalized power, for M. For formulation PF4 it is assumed that $M^* = M$.

Bounds on the system throughput at optimal generalized power for the formulation PF4 may be obtained by an argument similar to that of chapter 4 for ordinary power. Since $\rho_i^* = (\gamma^* R_i)/(\mu C_i)$ for $1 \leq i \leq M$, we may rewrite equation (5.85) in the form

$$\max_{1 \leq i \leq M} \frac{\gamma^* R_i}{\mu C_i} \geq \frac{r}{r+1} \geq \min_{1 \leq i \leq M} \frac{\gamma^* R_i}{\mu C_i}$$

This yields

$$\frac{r+1}{r\mu} \max_{1 \leq i \leq M} \frac{1}{C_i/R_i} \geq \frac{1}{\gamma^*} \geq \frac{r+1}{r\mu} \min_{1 \leq i \leq M} \frac{1}{C_i/R_i}$$

or

$$\frac{(r+1)/(r\mu)}{\min\limits_{1 \leq i \leq M} C_i/R_i} \geq \frac{1}{\gamma^*} \geq \frac{(r+1)/(r\mu)}{\min\limits_{1 \leq i \leq M} C_i/R_i}$$

Inverting this equation yields the following

Theorem 5.8

For the generalized power problem formulation PF4, the system throughput at maximal generalized power (the optimal traffic level) satisfies

$$\frac{r\mu}{r+1}\min_{1\le i\le M}\frac{C_i}{R_i} \le \gamma^* \le \frac{r\mu}{r+1}\max_{1\le i\le M}\frac{C_i}{R_i} \tag{5.86}$$

Setting $r=1$ in equations (5.84), (5.85), and (5.86) yield the bounds obtained in chapter 4 for the case of ordinary power.

If C_i/R_i are identical for all $1\le i\le M$, Theorem 5.8 shows that the optimal throughput γ^* is the common value

$$\gamma^* = \frac{r\mu C_i}{(r+1)R_i}$$

In this case, equations (5.85) and (5.84) give

$$\rho_i^* = \frac{r}{r+1}$$

and

$$\bar{N}_i^* = r$$

for all $1\le i\le M$. Note that Bharath-Kumar's results for the M/M/1 series network and the results of section 5.6.1 above for the M/M/1 parallel network are special instances of the theorems derived for PF4. We have $R_i=1$ for $1\le i\le M$ for the M/M/1 series network, while $R_i=p_i$ for $1\le i\le M$ for the M/M/1 parallel network (with known routing) in the above theorems.

5.8 The Queueing System G/M/1

In this section we briefly indicate how the analysis of generalized power P_r may be extended to the queueing system G/M/1. The analysis will follow closely that of chapter 2 for ordinary power, so we will simply sketch proofs of the results. As in chapter 2, we consider a G/M/1 system with mean interarrival time $\bar{t}$ and mean (exponential) service time $\bar{x}=1/\mu$. The average arrival rate to the system is $\lambda=1/\bar{t}$ and the throughput γ is equal to λ. The equilibrium probabilities are given in terms of σ, the root between 0 and 1 which satisfies

$$\sigma = \hat{A}(\mu-\mu\sigma)$$

where $\hat{A}$ is the Laplace transform for the interarrival time density. Recall that $\bar{N}=\rho/(1-\sigma)$ for G/M/1, and so $T=\bar{x}/(1-\sigma)$. In order to maximize generalized power P_r, we use equation (5.12). Differentiating T with respect to γ (as in equation (2.55)), we have

$$\frac{dT}{d\gamma} = \frac{T}{1-\sigma} \cdot \frac{d\sigma}{d\gamma}$$

and so, by equation (5.12),

$$rT^* = \frac{\gamma^* T^*}{1-\sigma^*} \cdot \left.\frac{d\sigma}{d\gamma}\right|_{\gamma=\gamma^*}$$

Using Little's result we have

$$rT^*(1-\sigma^*) = \overline{N}^* \cdot \left.\frac{d\sigma}{d\gamma}\right|_{\gamma=\gamma^*}$$

or

$$r\bar{x} = \overline{N}^* \cdot \left.\frac{d\sigma}{d\gamma}\right|_{\gamma=\gamma^*} \tag{5.87}$$

This equation relates various parameters at the maximum generalized power point for the queue G/M/1 and reduces to the ordinary power equation (2.56) when $r = 1$. Using equation (5.87), various interarrival time processes may be analyzed.

We first consider the queue $E_k/M/1$. Recall from equation (2.57) that σ satisfies

$$\sigma = \left(\frac{k\rho}{1-\sigma+k\rho}\right)^k$$

and so, from equation (2.58),

$$\frac{d\sigma}{d\gamma} = \frac{k\bar{x}\sigma(1-\sigma)}{\rho(1-\sigma+k\rho-k\sigma)}$$

Substituting this expression into equation (5.87), we have

$$r\bar{x} = \frac{\rho^*}{1-\sigma^*} \cdot \frac{k\bar{x}\sigma^*(1-\sigma^*)}{\rho^*(1-\sigma^*+k\rho^*-k\sigma^*)}$$

This yields

$$r(1-\sigma^*+k\rho^*-k\sigma^*) = k\sigma^*$$

or

$$r(1-\sigma^*+k\rho^*) = (r+1)k\sigma^*$$

Thus we have

$$\sigma^* = \frac{r(1+k\rho^*)}{(r+1)k+r} \tag{5.88}$$

and also

$$\rho^* = \frac{(r+1)k\sigma^* - r(1-\sigma^*)}{rk} \tag{5.89}$$

Note that equations (5.88) and (5.89) reduce to equations (2.59) and (2.60) for the case of ordinary power ($r = 1$). Substituting equation (5.89) into equation (2.57), we find that σ^* must satisfy

$$\sigma^* = \left[\frac{(r+1)k\sigma^* - r(1-\sigma^*)}{(r+1)k\sigma^*}\right]^k$$

which is equivalent to

$$\sigma^* = \left[1 - \frac{r(1-\sigma^*)/(r+1)\sigma^*}{k}\right]^k \tag{5.90}$$

This equation can be solved numerically to find the generalized power point for $E_k/M/1$.

If we let $k \to \infty$, we obtain results for the system $D/M/1$. In this case, equation (5.89) becomes

$$\rho^* = \frac{r+1}{r}\sigma^* \tag{5.91}$$

and equation (5.90) becomes

$$\sigma^* = e^{-[r(1-\sigma^*)/(r+1)\sigma^*]} \tag{5.92}$$

Using equation (5.87), we may extend our generalized power analysis to other $G/M/1$ systems in the same manner as was done for ordinary power in chapter 2. For example, certain $H_R/M/1$ systems may be studied. We choose not to explore this area further, but instead leave such details to the interested reader.

With the $G/M/1$ analysis, we conclude our examination of generalized power for simple network optimization problems. Although our analysis involved the generalized power function $P_r = P_{G,r} = \gamma'/T$, these results apply equally well if we use the generalized power version of Kleinrock's power function, namely $P_{K,r} = \rho'/(T/\bar{x})$. As noted above, this is because $\bar{x}$ is constant for all problems studied in this chapter. In the next chapter, we consider more general formulations (including the optimization problem of chapter 3) and find that the optimal solution does depend upon the particular generalized power function which is used. As in the case of ordinary power, we find that Kleinrock's generalized power extension yields optimal points with more pleasing properties than the generalized power function based on Giessler's definition.

CHAPTER 6
Generalized Power of the M/M/1 Parallel Network

In the previous chapter we studied a family of functions, indexed by a scalar parameter r, called generalized power which enabled the analyst to vary the importance of throughput relative to delay. The maximization of these functions was examined for various network topologies and optimization problem formulations. In each case, the problem involved optimizing over a single variable only, and thus (for any particular choice of r) the equation

$$rT = \gamma \frac{dT}{d\gamma}$$

could be used to obtain the optimal solution which maximized generalized power P_r. It was also the case that, for each of the formulations, the mean service time $\bar{x}$ was constant, and so the generalized power families which extended the three regular power functions (those of Giessler, Nakamura, and Kleinrock) all yielded the same optimal solution. Thus the function P_r of chapter 5 could refer to any of the generalized power definitions. However, if we study multiple variable optimization problems (such as the M/M/1 parallel network with unknown routing of chapter 3) in terms of generalized power, the optimal solution *will* depend on which power function we wish to extend. In section 6.1 of this chapter, we first examine the M/M/1 parallel network for the generalized power function based upon Giessler's power. The analysis is similar to that of chapter 3, and one result is that there are several negative aspects of this generalized power family. In section 6.2 we then examine this network for the generalized power extension due to Kleinrock (analogous to section 4.2.2), and we find that some of these undesirable properties disappear. Both sections follow the corresponding developments (of chapter 3 and section 4.2.2) very closely and in numerous cases are almost word for word; the new portions are those which come about due to the introduction of the parameter r. The quantity "2" which appears in several theorems of chapter 3 and section 4.2.2 is replaced in many of the corresponding theorems of this chapter by the quantity "$r+1$". Thus the reader who has studied the previous chapters of this dissertation may skip the proofs here and still obtain a sense of the results of this chapter.

6.1 Generalized Power (Extension of P_G)

Consider an M/M/1 parallel net with M channels which has no restriction on the routing, and optimize it with respect to the generalized power function

$$P_r = P_{G,r} = \frac{\gamma^r}{T} \tag{6.1}$$

Note that we have used the extension of Giessler's power function (which we obtain when $r = 1$). The analysis will parallel that of chapter 3; however many of the proofs for arbitrary r are more difficult than for the case $r = 1$ studied above. Therefore, we will include as much detail as possible to enable the reader to follow the presentation without leafing back and forth between this section and chapter 3. Recall that this parallel network may be used to represent either a single user with multiple paths for his packets or multiple users, each with a single path.

6.1.1 Description of the Optimization Problem

We consider the following multiple variable optimization problem. We are given M parallel channels with capacities $C_1, \ldots, C_M$, and we wish to find inputs $\lambda_1, \ldots, \lambda_M$ which maximize the generalized power of the system. The M variables of optimization are $\lambda_1, \ldots, \lambda_M$, and (as in chapter 3) the feasible region is the (compact) set of all vectors $(\lambda_1, \ldots, \lambda_M)$ such that

$$0 \leq \lambda_j \leq \mu C_j \qquad 1 \leq j \leq M$$

The objective function P_r of our optimization problem is given in equation (6.1). As in chapter 5, we assume that r is a positive real number, so that P_r is an increasing function of the throughput γ. Expressing P_r in terms of the variables of optimization λ_i, we obtain

$$P_r = \frac{\gamma^r}{T} = \frac{\gamma^r}{\sum_{i=1}^{M} \frac{\lambda_i}{\gamma} T_i} = \frac{\gamma^{r+1}}{\sum_{i=1}^{M} \lambda_i T_i}$$

or, since $\gamma = \lambda_1 + \cdots + \lambda_M$ (as in chapter 3),

$$P_r = \frac{\left(\sum_{i=1}^{M} \lambda_i\right)^{r+1}}{\sum_{i=1}^{M} \lambda_i T_i} \tag{6.2}$$

We wish to find a point which maximizes generalized power over the feasible region. Since $\rho_j = \lambda_j / \mu C_j$ (for all $1 \leq j \leq M$), this optimization problem is equivalent to one involving the M unknowns $\rho_1, \ldots, \rho_M$. We will frequently skip back and forth between these two equivalent formulations of this maximization problem. As in chapter 3, to find the global maximizer of the generalized power function P_r over the feasible region, we must find all critical points of P_r (points with $\nabla P_r = 0$) interior to the feasible region and compare them to the maximal boundary point.

8.1.2 Characterization of the Optimal Solution

Let us now examine the characteristics of the optimal boundary point of our optimization problem. We first renumber the channels of our parallel configuration so that $C_1 \geq C_2 \geq \cdots \geq C_M$. As was the case with regular power, we note that the optimal boundary point occurs when $\lambda_M = 0$ (the slowest channel has no throughput). This is a consequence of the following

Lemma: Consider two parallel M/M/1 systems, each with M channels, such that $C_{1i} \geq C_{2i}$ for $1 \leq i \leq M$ (that is, each channel of the first system has higher capacity than the corresponding channel of the second system). Then the optimal value of generalized power for the first system is at least as large as the optimal value of generalized power for the second system, i.e., $P^*_{r,1} \geq P^*_{r,2}$.

The proof of this lemma follows the corresponding proof for ordinary power in chapter 3. We first note that any point $(\lambda_1, \ldots, \lambda_M)$ which is feasible for the second system is also feasible for the first system. Also, for any such point feasible for the second system, we have

$$P_{r,1}(\lambda_1, \ldots, \lambda_M) = \frac{\left(\sum_{i=1}^{M} \lambda_i\right)^{r+1}}{\sum_{i=1}^{M} \lambda_i T_{1i}} \geq \frac{\left(\sum_{i=1}^{M} \lambda_i\right)^{r+1}}{\sum_{i=1}^{M} \lambda_i T_{2i}} = P_{r,2}(\lambda_1, \ldots, \lambda_M)$$

because (for M/M/1)

$$T_{1i} = \frac{1}{\mu C_{1i} - \lambda_i} \leq \frac{1}{\mu C_{2i} - \lambda_i} = T_{2i}$$

for $1 \leq i \leq M$. Therefore, if we let $(\lambda^*_{21}, \ldots, \lambda^*_{2M})$ be optimal for the second system, it is feasible for the first system, and we have

$$P^*_{r,2} = P_{r,2}(\lambda^*_{21}, \ldots, \lambda^*_{2M}) \leq P_{r,1}(\lambda^*_{21}, \ldots, \lambda^*_{2M}) \leq P^*_{r,1}$$

which gives the result. Applying the lemma to (pairs of) the M boundary faces $\lambda_i = 0$ justifies the above characterization of the optimal boundary point.

Thus (as in chapter 3) the optimal boundary point is obtained by considering a parallel system of $M - 1$ channels with capacities $C_1 \geq \cdots \geq C_{M-1}$ (the channel with the lowest capacity being dropped) and optimizing this system with respect to its corresponding generalized power function. Of course, the $M - 1$ channel system may have its optimal solution at a boundary point of its feasible region. To find such a point, channel $M - 1$ must be dropped, and a parallel system of $M - 2$ channels with capacities $C_1 \geq \cdots \geq C_{M-2}$ must be optimized. This process continues until the optimal solution for the M channel system is found. Note that a different generalized power function must be examined in each case, but the same value of the parameter r is, of course, always used.

Thus we see that the optimal solution of the M channel parallel system, which we denote by $(\lambda_1^*, \ldots, \lambda_M^*)$, satisfies

$$\lambda_i^* > 0 \qquad 1 \le i \le M^*$$
$$\lambda_i^* = 0 \qquad M^* < i$$

where the index M^* satisfies $1 \le M^* \le M$ (if $M^* = M$ then all components of the optimizer are positive and the solution is in the interior). The λ_i^* for $1 \le i \le M^*$ are found by considering a parallel system with M^* channels and capacities $C_1 \ge \cdots \ge C_{M^*}$ and determining the (interior) critical points for the generalized power function of this system. Therefore, solving the optimization problem for the M/M/1 parallel network is reduced to finding critical points of generalized power functions (with the same r) of the original network and certain subnetworks.

6.1.3 Determination of Critical Points

We now find points of an M/M/1 parallel system with M channels such that $\nabla P_r = 0$. Taking the partial derivative of P_r in equation (6.2) with respect to λ_j (where $1 \le j \le M$) yields

$$\frac{\partial P_r}{\partial \lambda_j} = \frac{(r+1)\left(\sum_{i=1}^{M} \lambda_i T_i\right)\left(\sum_{i=1}^{M} \lambda_i\right)^r - \left(\sum_{i=1}^{M} \lambda_i\right)^{r+1}\left(\lambda_j \frac{dT_j}{d\lambda_j} + T_j\right)}{\left(\sum_{i=1}^{M} \lambda_i T_i\right)^2} \tag{6.3}$$

or

$$\frac{\partial P_r}{\partial \lambda_j} = \frac{\left(\sum_{i=1}^{M} \lambda_i\right)^{r+1}}{\left(\sum_{i=1}^{M} \lambda_i T_i\right)^2}\left[\frac{(r+1)\sum_{i=1}^{M} \lambda_i T_i}{\sum_{i=1}^{M} \lambda_i} - \left(\lambda_j \frac{dT_j}{d\lambda_j} + T_j\right)\right] \tag{6.4}$$

If we assume that $\nabla P_r = 0$, then $\partial P_r / \partial \lambda_j = 0$ for $1 \le j \le M$, and thus we have

$$\lambda_j \frac{dT_j}{d\lambda_j} + T_j = \frac{(r+1)\sum_{i=1}^{M} \lambda_i T_i}{\sum_{i=1}^{M} \lambda_i} = \lambda_k \frac{dT_k}{d\lambda_k} + T_k \qquad \forall j, k \tag{6.5}$$

Although this equation holds for a G/G/1 parallel net, expressions for T_j and thus $dT_j/d\lambda_j$ are not known for the general case. Therefore, in the following we will specialize to the case M/M/1.

We now wish to rewrite equations (6.5) in terms of ρ_j, the equivalent variables of optimization. Since we assume each channel acts as an M/M/1 queueing system, we may proceed as in chapter 3 (see equation (3.5)) and write

$$\lambda_j \frac{dT_j}{d\lambda_j} + T_j = \frac{1}{\mu C_j} \cdot \frac{1}{(1-\rho_j)^2} \tag{6.6}$$

Equations (6.5) and (6.6) thus yield the following result which is identical to that derived for ordinary power, namely, that $\nabla P_r = 0$ implies

$$\frac{1}{\mu C_j} \cdot \frac{1}{(1-\rho_j)^2} = \frac{1}{\mu C_k} \cdot \frac{1}{(1-\rho_k)^2} \qquad \forall j,k \tag{6.7}$$

or

$$(1-\rho_j)^2 C_j = (1-\rho_k)^2 C_k \qquad \forall j,k \tag{6.8}$$

Taking square roots yields

$$(1-\rho_j)\sqrt{C_j/C_k} = 1-\rho_k \qquad \forall j,k \tag{6.9}$$

This gives us $M-1$ independent *linear* equations in the M unknowns $\rho_1, \ldots, \rho_M$, which are the same as those for the case $r=1$ in chapter 3. However, the Mth independent equation will depend on the particular value of r.

To determine this Mth equation, we proceed as in chapter 3. Equation (6.5) yields

$$\lambda_j \frac{dT_j}{d\lambda_j} + T_j = \frac{(r+1)\sum_{i=1}^{M} \lambda_i T_i}{\sum_{i=1}^{M} \lambda_i} \qquad 1 \le j \le M$$

and so, using equation (6.6) and Little's result, we have

$$\frac{1}{\mu C_j} \cdot \frac{1}{(1-\rho_j)^2} = \frac{(r+1)\sum_{i=1}^{M} \overline{N}_i}{\sum_{i=1}^{M} \lambda_i} \qquad 1 \le j \le M$$

Multiplying the jth equation by λ_j and recalling that $\rho_j = \lambda_j/\mu C_j$ yields

$$\frac{\rho_j}{(1-\rho_j)^2} = \lambda_j \cdot \frac{(r+1)\sum_{i=1}^{M} \overline{N}_i}{\sum_{i=1}^{M} \lambda_i} \qquad 1 \le j \le M$$

Summing these M equations gives

$$\sum_{i=1}^{M} \frac{\rho_i}{(1-\rho_i)^2} = (r+1)\sum_{i=1}^{M} \overline{N}_i \tag{6.10}$$

Since we are considering the ith channel as an M/M/1 system, then $\overline{N}_i = \rho_i/(1-\rho_i)$ and so

$$\sum_{i=1}^{M} \frac{\rho_i}{(1-\rho_i)^2} = (r+1) \sum_{i=1}^{M} \frac{\rho_i}{1-\rho_i}$$

We therefore have

$$\sum_{i=1}^{M} \frac{\rho_i - (r+1)\rho_i + (r+1)\rho_i^2}{(1-\rho_i)^2} = 0$$

or

$$\sum_{i=1}^{M} \frac{(r+1)\rho_i^2 - r\rho_i}{(1-\rho_i)^2} = 0$$

Multiplying the numerator and denominator of the ith term by C_i yields

$$\sum_{i=1}^{M} \frac{[(r+1)\rho_i^2 - r\rho_i]\, C_i}{(1-\rho_i)^2 C_i} = 0$$

By equation (6.8) we see that the denominators of all M terms in the above sum are identical. Multiplying both sides of the above equality by this common value finally gives our Mth independent equation

$$\sum_{i=1}^{M} [(r+1)\rho_i^2 - r\rho_i]\, C_i = 0 \tag{6.11}$$

This equation may also be written as

$$\sum_{i=1}^{M} \rho_i \left(\rho_i - \frac{r}{r+1}\right) C_i = 0 \tag{6.12}$$

Therefore, to find points $(\rho_1, \ldots, \rho_M)$ such that $\nabla P_r = 0$, we must solve M equations in M unknowns. We note that $M-1$ of the equations are linear (equation (6.9)), while the Mth equation is quadratic (equation (6.11)). We also note that the $M-1$ linear equations are identical for all r, while the Mth quadratic equation depends on r. Our solution strategy is to use the $M-1$ linear equations to eliminate $\rho_2, \ldots, \rho_M$ (i.e., express each in terms of ρ_1) and then use the Mth quadratic equation to solve for ρ_1. Thus we could have 0, 1, or 2 points which are critical points of P_r, depending on the given parameters (the channel capacities C_i).

6.1.4 Characteristics of Critical Points

Let us examine these M equations in greater detail to obtain some interesting characteristics of critical points of P_r (and thus of the optimal generalized power point, since it is a critical point of the generalized power function of a parallel subsystem with M^* channels). Proceeding as in chapter 3, equation (6.9) yields

$$\rho_1 \geq \rho_2 \geq \cdots \geq \rho_M \tag{6.13}$$

for any point of the M channel parallel system where the gradient of the generalized power function P_r is zero. Recalling the nature of the optimal generalized power point $(\rho_1^*, \ldots, \rho_M^*)$, we have

$$\rho_1^* \geq \rho_2^* \geq \cdots \geq \rho_M^* \tag{6.14}$$

and also

$$\begin{aligned} \rho_i^* &> 0 \quad 1 \leq i \leq M^* \\ \rho_i^* &= 0 \quad M^* < i \end{aligned}$$

where the ρ_i^* are determined by deleting all channels with index $i > M^*$ and solving for $\nabla P_r = 0$ of this reduced network.

We may also obtain a bound on the value of the utilization of the fastest channel at maximum generalized power by first examining the value of ρ_1 for a critical point of the generalized power function P_r. We note that, if $\rho_1 < r/(r+1)$, then by equation (6.13), $\rho_i < r/(r+1)$ for all $1 \leq i \leq M$. This is impossible by equation (6.12), and we thus conclude that

$$\rho_1 \geq \frac{r}{r+1} \tag{6.15}$$

for a critical point of P_r. Therefore, by the characterization of the optimal generalized power point, we have

$$\rho_1^* \geq \frac{r}{r+1} \tag{6.16}$$

Rewriting equation (6.16) as

$$\rho_1^* \geq r(1-\rho_1^*)$$

and using the fact that $\bar{N} = \rho/(1-\rho)$ for M/M/1 yields

$$\bar{N}_1^* \geq r \tag{6.17}$$

Other characteristics of the optimal generalized power point are consequences of results given in chapter 5 above. However, they also may be derived in a direct manner which parallels the development in chapter 3, and it is this latter approach that we choose to follow. Proceeding as in chapter 3, we see that equation (6.10) and the fact that $\rho_i/(1-\rho_i) = \bar{N}_i$ for an M/M/1 system yields

$$\sum_{i=1}^{M} \frac{\bar{N}_i}{1-\rho_i} = (r+1)\sum_{i=1}^{M} \bar{N}_i$$

or

$$\sum_{i=1}^{M} \bar{N}_i \left[(r+1) - \frac{1}{1-\rho_i} \right] = 0$$

This becomes

$$\sum_{i=1}^{M} \bar{N}_i \left[\frac{r - (r+1)\rho_i}{1-\rho_i} \right] = 0$$

or

$$\sum_{i=1}^{M} \bar{N}_i \left[r - \frac{\rho_i}{1-\rho_i} \right] = 0$$

Again, we recall that $\rho_i/(1-\rho_i) = \bar{N}_i$, and so

$$\sum_{i=1}^{M} \bar{N}_i (r - \bar{N}_i) = 0$$

or

$$r \sum_{i=1}^{M} \bar{N}_i = \sum_{i=1}^{M} (\bar{N}_i)^2$$

The above equation holds for critical points of the generalized power function of an M channel M/M/1 parallel network (with unknown routing). We may once again exploit the nature of the global optimizer of the generalized power function (it is a critical point for an M^* channel net) to write

$$r \sum_{i=1}^{M^*} \bar{N}_i^* = \sum_{i=1}^{M^*} (\bar{N}_i^*)^2 \tag{6.18}$$

In fact, since $\rho_i^* = 0$ (and thus $\bar{N}_i^* = 0$) for $i > M^*$, we may actually regard the summations in equation (6.18) as including terms from 1 to M, which then yields the following

Theorem 6.1

For the M/M/1 parallel network (with unknown routing), the average number in system at maximum generalized power satisfies

$$r\bar{N}^* = r \sum_{i=1}^{M} \bar{N}_i^* = \sum_{i=1}^{M} (\bar{N}_i^*)^2 \tag{6.19}$$

We observe that this result is actually a particular case of Theorem 5.5 above, since, as noted in chapter 4, the M/M/1 parallel network (with unknown routing) is an instance of formulation PF1.

In the same manner as was done previously for the simple nets of chapter 5, we may use equation (6.19) to yield the following

Theorem 6.2

For the M/M/1 parallel network (with unknown routing), the average number in system at maximum generalized power satisfies

$$r\sum_{i=1}^{M}(r - \bar{N}_i^*) = \sum_{i=1}^{M}(r - \bar{N}_i^*)^2 \tag{6.20}$$

Note that this dual equation is also a consequence of Theorem 5.6.

Equation (6.19) also yields the following

Theorem 6.3

For the M/M/1 parallel network (with unknown routing), the average number in system at maximum generalized power satisfies

$$\bar{N}^* \leq Mr \tag{6.21}$$

Note that this bound is a particular case of Theorem 5.7. Of course, we actually have the tighter bound

$$\bar{N}^* \leq M^* r \tag{6.22}$$

from equation (6.18) (any channel j assigned an input rate $\lambda_j^* = 0$ does not contribute to $\bar{N}^*$).

6.1.5 Solution of the Generalized Power Problem

Before solving the M independent equations, let us introduce the same simplifying notation as in chapter 3. We first define, for $1 \leq j \leq M$,

$$S_j \triangleq \sqrt{C_j/C_1} \tag{6.23}$$

Note that $1 = S_1 \geq S_2 \geq \cdots \geq S_M$. We also define, for $1 \leq j \leq M$,

$$\theta_j \triangleq 1 - \rho_j \tag{6.24}$$

Choosing $k = 1$ in equation (6.9), we have (for $1 \leq j \leq M$)

$$1 - \rho_1 = (1 - \rho_j)\sqrt{C_j/C_1}$$

which is, using our new terminology,

$$\theta_j = \frac{\theta_1}{S_j} \tag{6.25}$$

Arguing as in chapter 3, we also have

$$T_j = \frac{T_1}{S_j} \qquad 1 \le j \le M \tag{6.26}$$

which determines the average times at the channels in terms of the given capacities for critical points of P_r.

We now solve the M independent equations (in equations (6.9) and (6.11)) for $\rho_1, \ldots, \rho_M$. Equation (6.11) yields

$$\sum_{i=1}^{M} [(r+1)\rho_i - r]\rho_i C_i = 0$$

which becomes, using the definition of θ_i,

$$\sum_{i=1}^{M} [(r+1)(1-\theta_i) - r](1-\theta_i) C_i = 0$$

or

$$\sum_{i=1}^{M} [1-(r+1)\theta_i](1-\theta_i) C_i = 0$$

Dividing by C_1 and recalling that $C_i/C_1 = S_i^2$, we have

$$\sum_{i=1}^{M} [1-(r+1)\theta_i](1-\theta_i) S_i^2 = 0$$

or

$$\sum_{i=1}^{M} [1-(r+2)\theta_i + (r+1)\theta_i^2] S_i^2 = 0$$

Using equation (6.25), we obtain

$$\sum_{i=1}^{M} \left[1-(r+2)\frac{\theta_1}{S_i} + (r+1)\left(\frac{\theta_1}{S_i}\right)^2\right] S_i^2 = 0$$

or

$$\sum_{i=1}^{M} [(r+1)\theta_1^2 - (r+2)S_i\theta_1 + S_i^2] = 0$$

Having eliminated $\theta_2, \ldots, \theta_M$, we finally have the following quadratic equation in the unknown variable θ_1:

$$(r+1)M\theta_1^2 - (r+2)\left(\sum_{i=1}^{M} S_i\right)\theta_1 + \sum_{i=1}^{M} S_i^2 = 0 \tag{6.27}$$

The two roots of this equation are

$$u \triangleq \frac{(r+2)\sum_{i=1}^{M} S_i + \left[(r+2)^2\left(\sum_{i=1}^{M} S_i\right)^2 - 4(r+1)M\sum_{i=1}^{M} S_i^2\right]^{\frac{1}{2}}}{2(r+1)M} \tag{6.28}$$

and

$$v \triangleq \frac{(r+2)\sum_{i=1}^{M} S_i - \left[(r+2)^2\left(\sum_{i=1}^{M} S_i\right)^2 - 4(r+1)M\sum_{i=1}^{M} S_i^2\right]^{\frac{1}{2}}}{2(r+1)M} \tag{6.29}$$

These roots are real if and only if we have a non-negative discriminant, i.e.,

$$D_r \triangleq (r+2)^2\left(\sum_{i=1}^{M} S_i\right)^2 - 4(r+1)M\sum_{i=1}^{M} S_i^2 \geq 0 \tag{6.30}$$

In this case, we clearly have $0 < v \leq u$.

We now claim that

$$u \leq (r+1)v \tag{6.31}$$

Using equations (6.28) and (6.29), we see that equation (6.31) holds if and only if

$$(r+2)\sum_{i=1}^{M} S_i + \sqrt{D_r} \leq (r+1)(r+2)\sum_{i=1}^{M} S_i - (r+1)\sqrt{D_r}$$

This inequality is equivalent to

$$(r+2)\sqrt{D_r} \leq r(r+2)\sum_{i=1}^{M} S_i$$

or

$$\sqrt{D_r} \leq r\sum_{i=1}^{M} S_i$$

Squaring this last equation and using the definition of D_r given in equation (6.30), we obtain

$$(r+2)^2\left(\sum_{i=1}^{M} S_i\right)^2 - 4(r+1)M\sum_{i=1}^{M} S_i^2 \leq r^2\left(\sum_{i=1}^{M} S_i\right)^2$$

or

$$4(r+1)\left(\sum_{i=1}^{M} S_i\right)^2 \leq 4(r+1)M\sum_{i=1}^{M} S_i^2$$

Thus equation (6.31) is equivalent to

$$\left(\sum_{i=1}^{M} S_i\right)^2 \leq M\sum_{i=1}^{M} S_i^2 \tag{6.32}$$

Note that equation (6.32) is identical to equation (3.31) of chapter 3, which was proved using

the Cauchy-Schwarz inequality. This shows equation (6.31). Therefore, whenever the roots u and v are real (i.e., whenever equation (6.30) holds), we have the bounds

$$0 < v \leq u \leq (r+1)v \tag{6.33}$$

Assuming the root u is real, let us see when it leads to a feasible solution. We denote the point corresponding to u as $\rho(u)$ (or $\lambda(u)$ or $\theta(u)$, depending on the variable of interest), but suppress the dependence on the particular root for notational convenience when it is clear from the context. We call the root u feasible if the corresponding point $\rho(u)$ is feasible. The condition for feasibility is $0 \leq \rho_j < 1$ for $1 \leq j \leq M$. For the point $\theta(u)$, we have $\theta_1 = u$, and from equation (6.25), $\theta_j = u/S_j$ for $1 \leq j \leq M$. Equation (6.33) then gives $0 < \theta_j$ for all j, which is equivalent to $\rho_j < 1$. Thus the solution is feasible if and only if $\rho_j \geq 0$ for $1 \leq j \leq M$. This is equivalent to $\theta_j \leq 1$ for all j, or $u \leq S_j$. Since $S_j \geq S_k$ for $j \leq k$, the condition for feasibility becomes

$$u \leq S_M$$

(note that this is equivalent to $\rho_M \geq 0$). The feasible point is an interior point if and only if $u < S_M$ ($\rho_M > 0$). Similarly, the root v would give a feasible point if and only if

$$v \leq S_M$$

with the feasible point an interior point if and only if $v < S_M$. If these roots (u and v) are real and give rise to solutions within the feasible region, then they give values of θ_1 where the gradient of the generalized power function is zero. Using the relationship $\theta_j = \theta_1/S_j$ where $\theta_j = 1 - \rho_j$, all points $(\rho_1, \ldots, \rho_M)$ which satisfy $\nabla P_r = 0$ for the M channel parallel network may be found. As described above, solutions of this form (for the original network and certain subnetworks) will lead to the global maximizer of P_r over the region of feasibility.

Recall our assumption that both u and v are real (and thus equation (6.33) holds). If $S_M < v$, then neither root yields a feasible solution. If $v \leq S_M < u$, then only v yields a feasible solution, and we may consequently disregard u. We now show that if $u \leq S_M$ so that both u and v yield feasible solutions, then the solution corresponding to v gives higher generalized power P_r. Therefore, the root u need not be considered in determining optimal generalized power points. To this end, assume u and v are real and the solution points corresponding to them are both feasible. We examine the solution given by the root v. As in chapter 3, for $1 \leq i \leq M$ we have

$$\overline{N}_i = \frac{S_i - v}{v}$$

and

$$\lambda_i = \mu C_1 S_i (S_i - v)$$

Therefore, the value of generalized power at this feasible point (call the value $P_r(v)$) is

$$P_r(v) = \frac{\left(\sum_{i=1}^{M}\lambda_i\right)^{r+1}}{\sum_{i=1}^{M}\bar{N}_i} = \frac{(\mu C_1)^{r+1}\left[\sum_{i=1}^{M}S_i(S_i - v)\right]^{r+1}}{\frac{1}{v}\sum_{i=1}^{M}(S_i - v)}$$

or

$$P_r(v) = (\mu C_1)^{r+1} v \frac{\left[\sum_{i=1}^{M}S_i^2 - v\sum_{i=1}^{M}S_i\right]^{r+1}}{\sum_{i=1}^{M}S_i - Mv}$$

As in chapter 3, we now use the well-known result that if y and z are the two roots of the quadratic equation $az^2 + bz + c = 0$, then $y + z = -b/a$ and $y \cdot z = c/a$. Applying this to the roots u and v of equation (6.27), we find

$$\sum_{i=1}^{M}S_i = \frac{(r+1)M}{r+2}(u+v)$$

and

$$\sum_{i=1}^{M}S_i^2 = (r+1)M(u \cdot v)$$

Using these expressions in the equation for $P_r(v)$ gives

$$P_r(v) = (\mu C_1)^{r+1} v \frac{\left[(r+1)Muv - v\frac{(r+1)M}{r+2}(u+v)\right]^{r+1}}{\frac{(r+1)M}{r+2}(u+v) - Mv}$$

or

$$P_r(v) = (\mu C_1)^{r+1} v \frac{[(r+1)Mv]^{r+1}\left[u - \frac{1}{r+2}(u+v)\right]^{r+1}}{M\left[\frac{r+1}{r+2}(u+v) - v\right]}$$

Therefore

$$P_r(v) = (\mu C_1)^{r+1} v \frac{[(r+1)Mv]^{r+1}\left[\frac{r+1}{r+2}u - \frac{1}{r+2}v\right]^{r+1}}{M\left[\frac{r+1}{r+2}u - \frac{1}{r+2}v\right]}$$

or

$$P_r(v) = (\mu C_1)^{r+1} v \frac{[(r+1)Mv]^{r+1}}{M}\left[\frac{(r+1)u - v}{r+2}\right]^{r}$$

We finally have

$$P_r(v) = (\mu C_1)^{r+1} \frac{(r+1)^{r+1}}{(r+2)^r} M^r v^{r+2} |(r+1)u - v|^r \tag{6.34}$$

In a similar fashion, we examine the solution corresponding to u (call the value $P_r(u)$) and find that

$$P_r(u) = (\mu C_1)^{r+1} \frac{(r+1)^{r+1}}{(r+2)^r} M^r u^{r+2} |(r+1)v - u|^r \tag{6.35}$$

We claim that $P_r(v) \geq P_r(u)$. This is clearly true if $u = v$; equation (6.33) thus shows that we need only consider the range $u > v > 0$. Using equations (6.34) and (6.35), we see that we must prove

$$v^{r+2} |(r+1)u - v|^r \geq u^{r+2} |(r+1)v - u|^r$$

From equation (6.33) we observe that both sides of the above inequality are non-negative; therefore, we may take the rth root and preserve the inequality. This yields

$$v^{2/r} v |(r+1)u - v| \geq u^{2/r} u |(r+1)v - u|$$

or

$$[v^{1/r}]^2 v |(r+1)u - v| \geq [u^{1/r}]^2 u |(r+1)v - u| \tag{6.36}$$

We now prove equation (6.36). We first make the change of variable $a \triangleq u^{1/r}$ and $b \triangleq v^{1/r}$. Therefore, we have $a^r = u$ and $b^r = v$. Also note that $a > b > 0$ by the assumption that $u > v > 0$. Rewriting equation (6.36) in terms of these new variables, we have

$$b^{r+2} |(r+1)a^r - b^r| \geq a^{r+2} |(r+1)b^r - a^r|$$

or

$$a^{2r+2} - b^{2r+2} \geq (r+1)a^r b^r (a^2 - b^2)$$

This is equivalent to

$$[a^{r+1}]^2 - [b^{r+1}]^2 \geq (r+1)a^r b^r (a^2 - b^2)$$

or

$$(a^{r+1} + b^{r+1})(a^{r+1} - b^{r+1}) \geq (r+1)a^r b^r (a+b)(a-b) \tag{6.37}$$

We now note that $a^{r+1} - b^{r+1}$ may be factored as

$$a^{r+1} - b^{r+1} = (a-b) \sum_{k=0}^{r} a^k b^{r-k} = (a-b) \sum_{k=0}^{r} b^k a^{r-k} \tag{6.38}$$

Using equation (6.38) in equation (6.37) and then dividing the result by the positive term $(a-b)$, we have

$$a^{r+1}\sum_{k=0}^{r} a^k b^{r-k} + b^{r+1}\sum_{k=0}^{r} b^k a^{r-k} \geq (r+1)a^r b^r (a+b)$$

or

$$a^{r+1}b^r\sum_{k=0}^{r}\left(\frac{a}{b}\right)^k + b^{r+1}a^r\sum_{k=0}^{r}\left(\frac{b}{a}\right)^k \geq (r+1)a^r b^r (a+b)$$

Dividing both sides by the common positive factor $a^r b^r$ and rearranging terms yields

$$a\sum_{k=0}^{r}\left(\frac{a}{b}\right)^k - (r+1)a + b\sum_{k=0}^{r}\left(\frac{b}{a}\right)^k - (r+1)b \geq 0$$

or

$$a\sum_{k=0}^{r}\left[\left(\frac{a}{b}\right)^k - 1\right] + b\sum_{k=0}^{r}\left[\left(\frac{b}{a}\right)^k - 1\right] \geq 0$$

We therefore have

$$a\sum_{k=0}^{r}\left(\frac{a^k - b^k}{b^k}\right) - b\sum_{k=0}^{r}\left(\frac{a^k - b^k}{a^k}\right) \geq 0$$

or

$$\sum_{k=0}^{r}(a^k - b^k)\left(\frac{a}{b^k} - \frac{b}{a^k}\right) \geq 0$$

Thus we finally have

$$\sum_{k=0}^{r}(a^k - b^k)\left(\frac{a^{k+1} - b^{k+1}}{a^k b^k}\right) \geq 0 \qquad (6.39)$$

But equation (6.39) is obviously true since $a > b > 0$, and so equation (6.36) holds. Therefore $P_r(v) \geq P_r(u)$, and we consequently may disregard the root u in determining the optimal generalized power point.

Let us now review the solution procedure for finding the global maximizer of generalized power for an M/M/1 parallel network with M channels and channel capacities $C_1 \geq \cdots \geq C_M$. Let P_r^m $(1 \leq m \leq M)$ be the generalized power function for the subnetwork consisting of the fastest m channels (channel capacities $C_1, \ldots, C_m$), and let u_m and v_m be the corresponding roots when P_r^m is used. We first calculate $v = v_M$ from equation (6.29). If $v_M \leq S_M$, then v_M corresponds to a feasible point $\rho(v_M) = (\rho_1(v_M), \ldots, \rho_M(v_M))$, which is a critical point of the original generalized power function $P_r = P_r^M$. The coordinates of this point are given by $\rho_i(v_M) = 1 - v_M/S_i$ for $1 \leq i \leq M$, and the value of P_r^M at this point is given by equation (6.34) where $u = u_M$ satisfies equation (6.28). This procedure is then carried out for subnetworks obtained from the original network by dropping the slowest channel in each case.

For example, v_{M-1} (which corresponds to a critical point $\rho(v_{M-1})$ of P_r^{M-1}, the generalized power function for $M-1$ channels) and u_{M-1} are calculated next. Among these M points (corresponding to the roots v_j, $1 \le j \le M$, of the M subnetworks), that one which is feasible and has the highest generalized power is the global maximizer for our optimization problem. The following illustrates the procedure.

Step 0: Set $m = M$

Step 1: Calculate v_m (and u_m)

Step 2: Check if v_m real; if not, go to Step 5

Step 3: Check if v_m feasible; if not, go to Step 5

Step 4: Evaluate $P_r^m(v_m) = P_r^M(v_m)$

Step 5: Check if $m = 1$; if not, set $m = m-1$ and go to Step 1

Step 6: Determine the global optimum from the solutions calculated in Step 4

Thus, in general, we must search over M points, although later in this chapter we obtain conditions on the input parameters (the capacities C_i) which reduce the number of points that we need to examine.

6.1.5.1 Equal Capacity Case

We now consider several examples in greater detail. For our first (simple) example, let us assume that all channels have the same capacity C. In this case, we have $S_j = 1$ for all $1 \le j \le M$, so that the root $v = v_M$ becomes

$$v = \frac{(r+2)M - \sqrt{(r+2)^2M^2 - 4(r+1)M^2}}{2(r+1)M} = \frac{(r+2)M - rM}{2(r+1)M} = \frac{1}{r+1}$$

From equation (6.25) we must have $\theta_j = \theta_k$ for all j, k, so that the root v_M yields

$$\theta_j = \frac{1}{r+1} \qquad 1 \le j \le M$$

We also have $\rho_j = \rho_k$, $\overline{N}_j = \overline{N}_k$, and $\lambda_j = \lambda_k$ for all j, k. Thus

$$\rho_j = \frac{r}{r+1}, \quad \overline{N}_j = r, \quad \lambda_j = (\mu C)\frac{r}{r+1} \qquad 1 \le j \le M$$

and the solution given by v_M is $\rho(v_M) = (r/\lceil r+1 \rceil, \ldots, r/\lceil r+1 \rceil)$. Therefore (writing $P_r = P_r^M$)

$$P_r(v_M) = P_r^M(v_M) = \frac{\gamma^{r+1}}{\overline{N}} = \frac{(M\lambda_1)^{r+1}}{M\overline{N}_1} = \frac{\lfloor M\mu C r/(r+1) \rfloor^{r+1}}{Mr} = (\mu C)^{r+1}\frac{(Mr)^r}{(r+1)^{r+1}}$$

We see that this must be the global maximum as follows. For any $1 \leq m < M$, the root v_m clearly yields the solution $\rho(v_m) = (r/[r+1], \ldots, r/[r+1], 0, \ldots, 0)$ where the first m components are $r/[r+1]$ and the last $M - m$ are zero. This has generalized power

$$P_r(v_m) = P_r^M(v_m) = P_r^m(v_m) = (\mu C)^{r+1} \frac{(mr)^r}{(r+1)^{r+1}}$$

and so the solution corresponding to v_M is globally optimal. Thus we have $\overline{N}_j^* = r$, and so this equal capacity parallel channel case generalizes the result of Kleinrock [Klei79] given in equation (5.3) for the M/M/1 single-node system.

6.1.5.2 The Two Channel Case

The next case we shall consider is that of a network having only two parallel channels with capacities $C_1 \geq C_2$ (that is, the case $M = 2$). Following the procedure discussed at the end of section 6.1.5 (with $M = 2$), we first find the solution corresponding to v_2 (the case $m = 1$) and then find the solution corresponding to v_1 (the case $m = 1$). Note that v_2 (if it is real and feasible) yields the optimal critical point of $P_r = P_r^2$, while v_1 (which is always real and feasible) yields the optimal boundary point. For notational convenience, define $S \triangleq S_2$ $(= \sqrt{C_2/C_1})$; thus we have $0 < S \leq 1$. We look for the optimal point satisfying $\nabla P_r = 0$ (which is given by $v = v_2$) by examining equation (6.29) for this network. The root v is

$$v = \frac{(r+2)(1+S) - \sqrt{D_r}}{4(r+1)}$$

where

$$D_r = (r+2)^2(1+S)^2 - 8(r+1)(1+S^2) = (r^2+4r+4)(1+2S+S^2) - (8r+8)(1+S^2)$$

Thus we have

$$D_r = (r^2-4r-4)S^2 + (2r^2+8r+8)S + (r^2-4r-4)$$

and so

$$v = \frac{(r+2)(1+S) - \sqrt{(r^2-4r-4)S^2 + (2r^2+8r+8)S + (r^2-4r-4)}}{4(r+1)}$$

In order for this root to be real, we must have

$$D_r = (r^2-4r-4)S^2 + (2r^2+8r+8)S + (r^2-4r-4) \geq 0 \qquad (6.40)$$

If

$$r^2 - 4r - 4 \geq 0$$

then equation (6.40) holds for all values of the parameter S, and the roots u and v are real for all S. This is equivalent to $(r-2)^2 \geq 8$ or (since r is positive) to $r \geq 2 + 2\sqrt{2}$.

Next suppose that $0 < r < 2+2\sqrt{2}$ or that

$$0 < r^2 < 4r+4$$

In this case, the roots u and v are real for certain values of S, but are not real for other values of S. For the existence of real roots, equation (6.40) shows that we must have

$$(4+4r-r^2)S^2-(8+8r+2r^2)S+(4+4r-r^2) \leq 0$$

This is equivalent to

$$S^2-\left(\frac{8+8r+2r^2}{4+4r-r^2}\right)S+1 \leq 0$$

or

$$S^2-2\left(\frac{4+4r+r^2}{4+4r-r^2}\right)S+\left(\frac{4+4r+r^2}{4+4r-r^2}\right)^2 \leq \left(\frac{4+4r+r^2}{4+4r-r^2}\right)^2-1$$

Therefore we have

$$\left(S-\frac{4+4r+r^2}{4+4r-r^2}\right)^2 \leq \frac{(4+4r+r^2)^2-(4+4r-r^2)^2}{(4+4r-r^2)^2}$$

or

$$\left(S-\frac{4+4r+r^2}{4+4r-r^2}\right)^2 \leq \frac{4r^2(4+4r)}{(4+4r-r^2)^2}$$

We finally have

$$\left(S-\frac{4+4r+r^2}{4+4r-r^2}\right)^2 \leq \left(\frac{4r}{4+4r-r^2}\right)^2(1+r)$$

Thus the condition for v to be real becomes

$$-\frac{4r\sqrt{1+r}}{4+4r-r^2} \leq S-\frac{4+4r+r^2}{4+4r-r^2} \leq \frac{4r\sqrt{1+r}}{4+4r-r^2}$$

or

$$\frac{4+4r+r^2-4r\sqrt{1+r}}{4+4r-r^2} \leq S \leq \frac{4+4r+r^2+4r\sqrt{1+r}}{4+4r-r^2}$$

Note that the right-hand term in the above expression satisfies

$$\frac{4+4r+r^2+4r\sqrt{1+r}}{4+4r-r^2}=1+\frac{2r^2+4r\sqrt{1+r}}{4+4r-r^2}>1$$

for $0 < r^2 < 4r+4$. Since $S \leq 1$ by convention, our bounds become

$$\frac{4+4r+r^2-4r\sqrt{1+r}}{4+4r-r^2} \leq S \leq 1$$

Let us define

$$B(r) \triangleq \frac{4+4r+r^2-4r\sqrt{1+r}}{4+4r-r^2} \tag{6.41}$$

Note that

$$B(r) = \frac{\left[2\sqrt{1+r}-r\right]^2}{4+4r-r^2}$$

Recall that $0 < r^2 < 4r+4$ by assumption, and so taking square roots yields $0 < r < 2\sqrt{1+r}$. Thus we have $B(r) > 0$. We may also write

$$B(r) = 1 - \frac{4r\sqrt{1+r}-2r^2}{4+4r-r^2}$$

Since $0 < r < 2\sqrt{1+r}$ as shown above, multiplying this inequality by the positive quantity $2r$ yields $0 < 2r^2 < 4r\sqrt{1+r}$. Therefore, $B(r) < 1$.

Thus u and v are real for $0 < r^2 < 4r+4$ if and only if $B(r) \leq S \leq 1$ where $0 < B(r) < 1$, and $B(r)$ is defined in equation (6.41). If this condition is not satisfied, P_r has no critical point; hence the boundary point given by v_1 is globally optimal. For $r^2 \geq 4r+4$, the roots u and v are real for all S; we define $B(r) = 0$ for this range of r. In this case, P_r always has critical points, and these must be compared to the optimal boundary point to find the global solution. Using this (extended) definition of $B(r)$, we see that, for any r, the roots u and v are real if and only if

$$B(r) \leq S \leq 1 \tag{6.42}$$

Under the restrictions given in equation (6.42) on the parameter S, let us examine the root $v = v_2$ and determine if it leads to a feasible solution. We have $\rho_1 = 1 - v$, so that

$$\rho_1 = \frac{(3r+2)-(r+2)S+\sqrt{(r^2-4r-4)S^2+(2r^2+8r+8)S+(r^2-4r-4)}}{4(r+1)} \tag{6.43}$$

Now $\theta_2 = v/S$ and $\rho_2 = 1 - \theta_2$, so that

$$\rho_2 = \frac{(3r+2)S-(r+2)+\sqrt{(r^2-4r-4)S^2+(2r^2+8r+8)S+(r^2-4r-4)}}{4(r+1)S} \tag{6.44}$$

Recalling the discussion of feasibility after equation (6.33), we see that v yields a feasible point if and only if $\rho_2 \geq 0$. Clearly, ρ_2 is non-negative for $S \geq (r+2)/(3r+2)$, so we examine the range $S < (r+2)/(3r+2)$. Then, in order for $\rho_2 \geq 0$, we must have

$$\sqrt{(r^2-4r-4)S^2+(2r^2+8r+8)S+(r^2-4r-4)} \geq (r+2)-(3r+2)S$$

Since both sides of this inequality are nonnegative in the range of S under consideration, squaring will preserve it, and our condition becomes

$$(r^2-4r-4)S^2+(2r^2+8r+8)S+(r^2-4r-4) \geq$$

$$(r^2+4r+4)-(6r^2+16r+8)S+(9r^2+12r+4)S^2$$

or

$$(8r+8)-(8r^2+24r+16)S+(8r^2+16r+8)S^2 \leq 0$$

Dividing by the positive term $8r+8$ yields

$$1-(r+2)S+(r+1)S^2 \leq 0$$

or

$$(1-S)[1-(r+1)S] \leq 0$$

Since the range under consideration is $S < (r+2)/(3r+2) < 1$, the above inequality holds if and only if $1-(r+1)S \leq 0$ or

$$\frac{1}{r+1} \leq S$$

Note that

$$\frac{1}{r+1} < \frac{r+2}{3r+2}$$

because

$$3r+2 < (r+1)(r+2) = r^2+3r+2$$

for $r > 0$. Putting together the above two cases, the condition for v to be feasible is

$$\frac{1}{r+1} \leq S$$

under the assumption that the roots are real.

We still must show that

$$B(r) < \frac{1}{r+1} \tag{6.45}$$

For $r^2 \geq 4r+4$, this is trivially true since we have defined $B(r) = 0$ for that range of r. For $0 < r^2 < 4r+4$, using equation (6.41), we see that equation (6.45) holds if and only if

$$\frac{4+4r+r^2-4r\sqrt{1+r}}{4+4r-r^2} < \frac{1}{r+1}$$

or

$$4(1+r)^2+r^2(1+r)-4r(1+r)\sqrt{1+r} < 4(1+r)-r^2$$

This is equivalent to

$$4r(1+r)+r^2(1+r)+r^2 < 4r(1+r)\sqrt{1+r}$$

or, dividing by the positive number r,

$$(4+r)(1+r)+r < 4(1+r)\sqrt{1+r}$$

Squaring this last inequality yields

$$(16+8r+r^2)(1+2r+r^2)+2r(4+5r+r^2)+r^2 < 16(1+2r+r^2)(1+r)$$

or

$$16+48r+44r^2+12r^3+r^4 < 16+48r+48r^2+16r^3$$

This simplifies to

$$r^4 < 4r^2+4r^3$$

or, dividing by the positive quantity r^2,

$$r^2 < 4+4r$$

which is true for this range of r. Thus, in all cases, equation (6.45) holds. This shows that the root $v = v_2$ is real and yields a feasible point (ρ_1, ρ_2) which is a critical point of P_r if and only if S satisfies

$$\frac{1}{r+1} \leq S \leq 1$$

The values for ρ_1 and ρ_2 are given above in equations (6.43) and (6.44).

We next consider the case $m = 1$ of our optimization procedure. To find the optimal boundary point (given by v_1), we must examine the face $\lambda_2 = 0$ (the slowest channel is dropped). This face corresponds to a single M/M/1 system with capacity C_1, and so, from the above-mentioned result of Kleinrock (see also equation (5.4)), we find that the optimal boundary point is $(\rho_1, \rho_2) = (r/[r+1], 0)$.

We now determine the optimal generalized power point (step 6 of our procedure), which depends on the given parameter S. Since an optimal interior point must have $\nabla P_r = 0$ (and thus corresponds to $v = v_2$), we immediately see that the global maximum occurs at the optimal boundary point when $S < 1/(r+1)$ (i.e., $C_1 > (r+1)^2 C_2$). For $S = 1/(r+1)$ (i.e., $C_1 = (r+1)^2 C_2$), we find that $v = S = 1/(r+1)$ as follows. First recall that v is given by

$$v = \frac{(r+2)(1+S) - \sqrt{(r+2)^2(1+S)^2 - 8(r+1)(1+S^2)}}{4(r+1)}$$

Now $r+1 = 1/S$, and so $r+2 = (1+S)/S$. Using these values we find

$$v = \frac{(1+S)^2/S - \sqrt{(1+S)^4/S^2 - 8(1+S^2)/S}}{4/S}$$

or

$$v = \frac{(1+S)^2 - \sqrt{(1+S)^4 - 8S(1+S^2)}}{4}$$

However, a bit of multiplication shows that $(1+S)^4 - 8S(1+S^2) = (1-S)^4$, and so

$$v = \frac{(1+S)^2 - (1-S)^2}{4} = \frac{4S}{4} = S$$

or

$$v = \frac{1}{r+1}$$

as desired. Thus $\rho_1 = 1 - v = r/(r+1)$ and $\rho_2 = 1 - (v/S) = 0$. In this case, the optimal boundary point and the critical point given by v coincide, and so this point must be the global maximum. Thus we have $(\rho_1^*, \rho_2^*) = (r/[r+1], 0)$ when $S \leq 1/(r+1)$ $(C_1 \geq (r+1)^2 C_2)$.

For $1/(r+1) < S \leq 1$ $(C_2 \leq C_1 < (r+1)^2 C_2)$, we must compare the value of generalized power at the point given by the root v satisfying $\nabla P_r = 0$ (which is an interior point for this range of S) with that from the optimal boundary point. Rather than finding the generalized power itself (a difficult computation), we will make use of the optimization theory fact stated in chapter 3 at the end of section 3.2. In order to determine the global maximizer for the range $1/(r+1) < S \leq 1$, we examine the gradient of the generalized power function at the optimal boundary point. Recall from equation (6.4) that (for $j = 1, 2$)

$$\frac{\partial P_r}{\partial \lambda_j} = \frac{(\lambda_1+\lambda_2)^{r+1}}{(\lambda_1 T_1 + \lambda_2 T_2)^2}\left[\frac{(r+1)(\lambda_1 T_1 + \lambda_2 T_2)}{\lambda_1+\lambda_2} - \left(\lambda_j \frac{dT_j}{d\lambda_j} + T_j\right)\right]$$

Using the fact that $dT/d\lambda = T^2$ for M/M/1 along with Little's result, the jth partial derivative simplifies to

$$\frac{\partial P_r}{\partial \lambda_j} = \frac{(\lambda_1+\lambda_2)^{r+1}}{(\overline{N}_1+\overline{N}_2)^2}\left[\frac{(r+1)(\overline{N}_1+\overline{N}_2)}{\lambda_1+\lambda_2} - T_j(\overline{N}_j+1)\right]$$

At $(\rho_1, \rho_2) = (r/[r+1], 0)$ we have $(\lambda_1, \lambda_2) = (r\mu C_1/[r+1], 0)$. At this optimal boundary point we also have $\overline{N}_1 = r$, $\overline{N}_2 = 0$, $T_1 = r/\lambda_1 = (r+1)/\mu C_1$, and $T_2 = 1/\mu C_2 = 1/S^2 \mu C_1$ (by the definition of S). Therefore,

$$\frac{\partial P_r}{\partial \lambda_1} = \frac{\left[\mu C_1 \frac{r}{r+1}\right]^{r+1}}{r^2}\left[\frac{(r+1)r}{r\mu C_1/(r+1)} - \frac{r+1}{\mu C_1}(r+1)\right] = 0$$

which is as expected. Also,

$$\frac{\partial P_r}{\partial \lambda_2} = \frac{\left[\mu C_1 \frac{r}{r+1}\right]^{r+1}}{r^2}\left[\frac{(r+1)^2}{\mu C_1} - \frac{1}{S^2 \mu C_1} \cdot 1\right]$$

or

$$\frac{\partial P_r}{\partial \lambda_2} = (\mu C_1)^r \left(\frac{r}{r+1}\right)^{r-1}\left[1 - \frac{1}{(r+1)^2 S^2}\right]$$

Thus, at the optimal boundary point $(\lambda_1, \lambda_2) = (r\mu C_1/[r+1], 0)$, we find that the gradient of P_r has value

$$\nabla P_r = (0, \mu C_1[1 - 1/(r+1)^2 S^2])$$

By the well-known result from optimization theory quoted in chapter 3, if a point is a local maximum of P_r, then $\nabla P_r \cdot \mathbf{d} \leq 0$ for every feasible direction $\mathbf{d}$. Clearly, for $1/(r+1) < S \leq 1$ the point $(r\mu C_1/[r+1], 0)$ cannot be even a local maximum because there are (infinitely many) feasible directions $\mathbf{d}$ with $\nabla P_r \cdot \mathbf{d} > 0$. In fact, at that point any vector $\mathbf{d} = (d_1, d_2)$ with $d_2 > 0$ is a feasible direction such that

$$\nabla P_r \cdot \mathbf{d} = (\mu C_1)^r \left(\frac{r}{r+1}\right)^{r-1}\left[1 - \frac{1}{(r+1)^2 S^2}\right] d_2 > 0$$

Thus for $1/(r+1) < S \leq 1$ the maximal boundary point is not globally optimal (it is not even locally optimal), and so the critical point given by the root v must be optimal. Writing S in terms of the given channel capacities, the above cases enable us to prove the following

Theorem 6.4

The optimal solution which maximizes the generalized power of the two channel M/M/1 parallel network is:

(a) *for* $C_2 \leq C_1 \leq (r+1)^2 C_2$ *then*

$$\rho_1^* = \frac{(3r+2) - (r+2)S + \sqrt{(r^2-4r-4)S^2 + (2r^2+8r+8)S + (r^2-4r-4)}}{4(r+1)} \tag{6.46}$$

$$\rho_2^* = \frac{(3r+2)S - (r+2) + \sqrt{(r^2-4r-4)S^2 + (2r^2+8r+8)S + (r^2-4r-4)}}{4(r+1)S} \tag{6.47}$$

(b) *for* $(r+1)^2 C_2 \leq C_1$ *then*

$$\rho_1^* = \frac{r}{r+1}, \quad \rho_2^* = 0 \tag{6.48}$$

We now examine the behavior of the optimal solution for case (a) of Theorem 6.4 in greater detail; in particular, we focus on the parameter ρ_1^*. Proceeding as in chapter 3, we first recall that

$$v = \frac{(r+2)(1+S) - \sqrt{D_r}}{4(r+1)}$$

where

$$D_r = (r^2 - 4r - 4)S^2 + (2r^2 + 8r + 8)S + (r^2 - 4r - 4)$$

for the range of S under consideration $(1/|r+1| \leq S \leq 1)$. Differentiating v with respect to S yields

$$\frac{dv}{dS} = \frac{1}{4(r+1)}\left[(r+2) - \frac{(r^2 - 4r - 4)S + (r^2 + 4r + 4)}{\sqrt{D_r}}\right]$$

and (after a bit of manipulation)

$$\frac{d^2v}{dS^2} = \frac{4r^2}{D^{3/2}}$$

Since $D_r > 0$ for $1/(r+1) \leq S \leq 1$, we have $d^2v/dS^2 > 0$, and thus v is a strictly convex function of S for this range of S. Therefore, $\rho_1^* = 1 - v$ is a strictly concave function of S. To find the maximum value of ρ_1^*, we set $d\rho_1^*/dS = 0$, or equivalently $dv/dS = 0$. This gives

$$(r+2)\sqrt{D_r} = (r^2 - 4r - 4)S + (r^2 + 4r + 4)$$

Squaring this equation and using the definition of D_r yields (after some calculation)

$$(r^2 - 4r - 4)S^2 + 2(r^2 + 4r + 4)S - (r^2 + 4r + 4) = 0$$

Note that the corresponding value of D_r satisfies

$$D_r = 2r^2$$

If $r^2 - 4r - 4 = 0$ $(r = 2 + \sqrt{2})$, we obtain

$$S = \frac{1}{2}$$

We also have that the maximum value of ρ_1^* is

$$\rho_1^* = \frac{2 + \sqrt{2}}{4}$$

If $r^2 - 4r - 4 > 0$ $(r > 2 + \sqrt{2})$, it can be shown that

$$S = \frac{(r+2)\left|r\sqrt{2} - (r+2)\right|}{r^2 - 4r - 4}$$

and that

$$\frac{1}{r+1} < S < 1$$

In this case, we also find $S < 1/2$, and the maximum value of ρ_1^* is

$$\rho_1^* = \frac{r^2 - r(2+\sqrt{2})}{r^2 - 4r - 4}$$

If $r^2 - 4r - 4 < 0$ $(r < 2+\sqrt{2})$, it can be shown that

$$S = \frac{(r+2)[(r+2) - r\sqrt{2}]}{4+4r-r^2}$$

and that

$$\frac{1}{r+1} < S < 1$$

In this case, we also find $S > 1/2$, and the maximum value of ρ_1^* is

$$\rho_1^* = \frac{r(2+\sqrt{2}) - r^2}{4+4r-r^2}$$

Several interesting consequences of this simple two channel parallel network may be drawn from the above results. Let us regard the network as a model of two users (source-destination pairs), each with its own channel for its packets. We observe that operating at the optimal generalized power point may be unfair the sense that some users are restricted to having zero throughput. In fact, from the above characterization of the global maximum of this system, we see that user 2 will have zero throughput whenever $(r+1)^2 C_2 \leq C_1$ (i.e., whenever the faster channel is at least $(r+1)^2$ times as fast as the slower channel). Of course, such a system operating point is unfair to user 2. Note that, as r increases, the set of capacities which yield fair optimal solutions also becomes larger. This is not surprising, because throughput is favored for large r. We also note that local generalized power is not necessarily equal to global generalized power, i.e., an operating point obtained by maximizing generalized power using only local information (where each user is aware only of traffic characteristics along his own path) may not be the operating point obtained by globally maximizing generalized power. We see that the optimal point using local information is $(\rho_1, \rho_2) = (r/\lceil r+1 \rceil, r/\lceil r+1 \rceil)$, which is globally optimal if and only if $S = 1$ (i.e., if and only if both channels are the same speed). Thus, an algorithm which attempts to optimize generalized power using only local information will, in general, fail.

6.1.6 Simplifying the Determination of the Optimal Solution

In the two channel example analyzed above, the global optimum was determined by examining the gradient of the generalized power function at the optimal boundary point. Using this example as a guide, we find that it is sometimes possible to analyticaliy determine the optimal solution without evaluating all the M roots v_j for $1 \leq j \leq M$. That is, we may eliminate the tail of our optimization procedure by stopping with root v_m for some index m.

Before we consider the general case, let us study the special case where only one root needs to be evaluated. We wish to find conditions on the parameters S_i (and thus the channel capacities C_i) which insure that the points corresponding to the roots $v_1, v_2, \ldots, v_{M-1}$ are not optimal. In such a case, our optimization procedure involves one iteration, and since the conditions we seek are in terms of the given S_i only, the remaining $M-1$ roots need not be evaluated. We proceed by examining the gradient of the generalized power function for the original M channel network at these $M-1$ points. In particular, consider an index $1 \leq m < M$ (which corresponds to a subnetwork with $m < M$ channels) and assume that v_m yields a feasible point which is a critical point of its generalized power function P_r^m. That is, v_m is real and satisfies $v_m \leq S_m$. The point $(\rho_1, \ldots, \rho_m)$ (or $(\lambda_1, \ldots, \lambda_m)$) corresponding to v_m is then given by $\rho_j = \rho_j(v_m) = 1 - v_m/S_j$ for $1 \leq j \leq m$ and satisfies

$$\lambda_j \frac{dT_j}{d\lambda_j} + T_j = \frac{(r+1)\sum_{i=1}^{m} \lambda_i T_i}{\sum_{i=1}^{m} \lambda_i} \qquad 1 \leq j \leq m \tag{6.49}$$

We now evaluate the gradient of $P_r = P_r^M$ at the point $(\lambda_1, \ldots, \lambda_m, 0, \ldots, 0)$. Note that this point represents the solution corresponding to the root v_m in the original M-dimensional space. Evaluating equation (6.4) for the jth partial derivative of $P_r = P_r^M$ at the above-mentioned point, we find (since $\lambda_j = 0$ for $m < j \leq M$)

$$\frac{\partial P_r}{\partial \lambda_j} = \frac{\left(\sum_{i=1}^{m} \lambda_i\right)^{r+1}}{\left(\sum_{i=1}^{m} \lambda_i T_i\right)^2} \left[\frac{(r+1)\sum_{i=1}^{m} \lambda_i T_i}{\sum_{i=1}^{m} \lambda_i} - \left(\lambda_j \frac{dT_j}{d\lambda_j} + T_j\right) \right]$$

From equation (6.49) we have

$$\lambda_1 \frac{dT_1}{d\lambda_1} + T_1 = \frac{(r+1)\sum_{i=1}^{m} \lambda_i T_i}{\sum_{i=1}^{m} \lambda_i} = \lambda_j \frac{dT_j}{d\lambda_j} + T_j \qquad 1 \leq j \leq m$$

and so (for $1 \leq j \leq M$)

$$\frac{\partial P_r}{\partial \lambda_j} = \frac{\left(\sum_{i=1}^{m} \lambda_i\right)^{r+1}}{\left(\sum_{i=1}^{m} \lambda_i T_i\right)^2} \left[\left(\lambda_1 \frac{dT_1}{d\lambda_1} + T_1\right) - \left(\lambda_j \frac{dT_j}{d\lambda_j} + T_j\right)\right]$$

Thus, we clearly have

$$\frac{\partial P_r}{\partial \lambda_j} = 0 \qquad 1 \leq j \leq m$$

as expected.

We now examine the jth partial derivative for $m < j \leq M$. Using equation (6.6) and the fact that $\lambda_j = 0$ in the range of j under consideration, we find

$$\frac{\partial P_r}{\partial \lambda_j} = \frac{\left(\sum_{i=1}^{m} \lambda_i\right)^{r+1}}{\left(\sum_{i=1}^{m} \lambda_i T_i\right)^2} \left[\frac{1}{\mu C_1} \cdot \frac{1}{(1-\rho_1)^2} - T_j\right]$$

Now (as $\lambda_j = 0$),

$$T_j = \frac{1}{\mu C_j} = \frac{1}{\mu C_1} \cdot \frac{C_1}{C_j} = \frac{1}{\mu C_1} \cdot \frac{1}{S_j^2}$$

and so

$$\frac{\partial P_r}{\partial \lambda_j} = \frac{\left(\sum_{i=1}^{m} \lambda_i\right)^{r+1}}{\left(\sum_{i=1}^{m} \lambda_i T_i\right)^2} \left[\frac{1}{\mu C_1} \cdot \frac{1}{(1-\rho_1)^2} - \frac{1}{\mu C_1} \cdot \frac{1}{S_j^2}\right]$$

Therefore,

$$\frac{\partial P_r}{\partial \lambda_j} = \frac{1}{\mu C_1} \cdot \frac{1}{(1-\rho_1)^2} \frac{\left(\sum_{i=1}^{m} \lambda_i\right)^{r+1}}{\left(\sum_{i=1}^{m} \lambda_i T_i\right)^2} \left[1 - \left(\frac{1-\rho_1}{S_j}\right)^2\right]$$

or, using $\rho_1 = \rho_1(v_m) = 1 - v_m$,

$$\frac{\partial P_r}{\partial \lambda_j} = \frac{1}{\mu C_1} \cdot \frac{1}{(v_m)^2} \cdot \frac{\left(\sum_{i=1}^{m} \lambda_i\right)^{r+1}}{\left(\sum_{i=1}^{m} \lambda_i T_i\right)^2} \left[1 - \left(\frac{v_m}{S_j}\right)^2\right]$$

for $m < j \leq M$. Thus we have

$$\frac{\partial P_r}{\partial \lambda_j} = \begin{cases} 0 & 1 \le j \le m \\ K\left[1 - \left(\frac{v_m}{S_j}\right)^2\right] & m < j \le M \end{cases} \tag{6.50}$$

where $K > 0$.

We will now use the above evaluation of the jth partial derivative to find conditions on the given parameters S_i which insure that the solution corresponding to the root v_m is not the global optimum. To this end, we evaluate $\nabla P_r \cdot \mathbf{d}$ at the point given by v_m where $\mathbf{d}$ is any feasible direction. Such a feasible direction must be of the form $\mathbf{d} = (d_1, \ldots, d_M)$ where $d_j > 0$ for $m < j \le M$. We have

$$\nabla P_r \cdot \mathbf{d} = \sum_{j=1}^{M} \frac{\partial P_r}{\partial \lambda_j} d_j$$

which, using equation (6.50), becomes

$$\nabla P_r \cdot \mathbf{d} = K \sum_{j=m+1}^{M} d_j \left[1 - \left(\frac{v_m}{S_j}\right)^2\right]$$

One condition that insures $\nabla P_r \cdot \mathbf{d} > 0$ (and, therefore, that the solution corresponding to v_m is not optimal) is to have

$$1 - \left(\frac{v_m}{S_j}\right)^2 > 0 \qquad m < j \le M$$

or

$$v_m < S_j \qquad m < j \le M$$

Since $S_M \le S_j$ for all $1 \le j \le M$, this condition becomes

$$v_m < S_M$$

Therefore, if the root v_m satisfies $v_m < S_M$, then the solution point corresponding to v_m is not even a local maximum, because there are feasible directions $\mathbf{d}$ such that $\nabla P_r \cdot \mathbf{d} > 0$ at that point.

Note that the verification of the condition $v_m < S_M$ involves the calculation of the root v_m, which is precisely what we are trying to avoid. Thus we wish to find such a condition which depends on the input parameters S_i only. Recall from equation (6.15) that we have $\rho_1 \ge r/(r+1)$ for a critical point $(\rho_1, \ldots, \rho_m)$ of P_r^m. Then we must have $v_m \le 1/(r+1)$ since $v_m = 1 - \rho_1$. Hence, for those systems for which $S_M > 1/(r+1)$, we have $v_m < S_M$, and thus, by the above argument, the solution point corresponding to v_m is not optimal. Since this is true for any $m < M$, the point corresponding to the root v_M must be globally optimal. We have proved the following

Theorem 6.5

If $S_M > 1/(r+1)$ (i.e., $C_1/C_M < (r+1)^2$), then the global maximum of P_r for the M channel parallel network is given by the point which corresponds to the root v_M.

This yields the promised condition based on the given parameters S_i which guarantees that only one solution point (that corresponding to v_M) needs to be calculated in determining the optimal generalized power point. Note that the result for the two channel example in the range $1/(r+1) < S \leq 1$ follows from the above theorem.

Now let us return to the problem of finding simplifying conditions for the general case. The procedure which led to the proof of Theorem 6.5 may be generalized in the following way. Define n as the largest index ($1 \leq n \leq M$) for which $S_n > 1/(r+1)$. Therefore, $S_i > 1/(r+1)$ (i.e., $C_i/C_1 > 1/(r+1)^2$) for $1 \leq i \leq n$, and $S_i \leq 1/(r+1)$ (i.e., $C_i/C_1 \leq 1/(r+1)^2$) for $i > n$. Note that Theorem 6.5 is the case $n = M$. Thus the ratio of the capacity of the fastest channel to that of any channel up to and including channel n is less than $(r+1)^2$ ($C_1/C_i < (r+1)^2$ for $1 \leq i \leq n$), while its ratio to the capacities of all channels strictly slower than channel n is at least $(r+1)^2$ ($C_1/C_i \geq (r+1)^2$ for $n < i$). Applying the previous argument to the subnetwork with n channels, we see that the solutions corresponding to roots $v_{n-1}, \ldots, v_1$ must yield lower generalized power than the solution corresponding to v_n. Therefore, the optimal generalized power point must be given by one of the roots $v_M, \ldots, v_n$, and we may ignore the others. We have shown the following

Theorem 6.6

Let n be the largest index ($1 \leq n \leq M$) such that $S_n > 1/(r+1)$. The global maximum of P_r for the M channel parallel network is given by a point which corresponds to one of the roots $v_M, \ldots, v_n$.

This enables us to greatly simplify the determination of the optimal generalized power point for certain networks. In terms of the procedure given at the end of section 6.1.5, we may insert the following new step between steps 4 and 5:

Step 4.5: Check if $S_m > 1/(r+1)$; if yes, go to Step 6

Thus as soon as $S_m > 1/(r+1)$, we may immediately disregard any subsequent roots. That is, we have found a condition which is easy to check and which (if true) enables us to eliminate the tail of our optimization procedure.

6.1.7 Fairness

We now demonstrate that the above two theorems yield results concerning fairness (as defined in chapter 3). We first consider Theorem 6.5. Note that the optimal point (given by v_M) is an interior point of the feasible region since $v_M \leq 1/(r+1) < S_M$, and so it yields a fair solution point. We have shown the following

Theorem 6.7

If $S_M > 1/(r+1)$ (i.e., $C_1/C_M < (r+1)^2$), then the global maximum of P_r for the M channel parallel network is a fair operating point.

Thus we see that if the capacity of the fastest channel is less than $(r+1)^2$ times the capacity of the slowest channel, the optimal operating point with respect to generalized power gives each channel (or user) non-zero throughput. Therefore, an unfair optimal solution can only occur when the ratio of the fastest channel capacity to the slowest is at least $(r+1)^2$. However, as we shall see in section 6.1.7.1, this property does not characterize unfair solutions. That is, there are parallel networks satisfying $C_1/C_M \geq (r+1)^2$ which yield fair global maxima (i.e., each user has non-zero throughput).

The above result that $S_M > 1/(r+1)$ implies the optimal solution is fair (which used Theorem 6.5) may be generalized using Theorem 6.6 and the following claim.

Claim: If $1 \leq i \leq j \leq M$ and $S_i > 1/(r+1)$, then the solution corresponding to root v_j (for the subnetwork with j channels) is fair for user i.

We prove this claim by first noting that $v_j \leq 1/(r+1) < S_i$, and also (since $i \leq j$, and thus the subnetwork with j channels necessarily includes channel i) that $\rho_i(v_j) = 1 - v_j/S_i$. Therefore $\rho_i > 0$, and the solution is fair for user i. Thus the claim holds, which enables us to prove the following

Theorem 6.8

If for some m $(1 \leq m \leq M)$ $S_m > 1/(r+1)$ (i.e., $C_1/C_m < (r+1)^2$), then the global maximum of P_r for the M channel parallel network is fair for user m in the sense that channel m is assigned non-zero throughput at the optimal generalized power point.

To prove this, we first define n as in Theorem 6.6; that is, we let n be the largest index $(1 \leq n \leq M)$ such that $S_n > 1/(r+1)$. Thus we must have $1 \leq m \leq n$. From Theorem 6.6, we know that the optimal generalized power point corresponds to one of the roots $v_M, \ldots, v_n$. We now note that this optimal point will be fair for users 1 through n (and thus for user m) no matter which of the roots $v_M, \ldots, v_n$ yields highest generalized power. This is true from the

claim proved above, because if we let v_j yield the optimal point and i be one of these first n users, then $i \leq n \leq j$. Thus root v_j yields a point which gives non-zero throughput to each of the channels (users) $1, \ldots, n$ and therefore to user m, which proves Theorem 6.8

This shows that if the ratio of the capacity of the fastest channel to the capacity of channel m $(1 \leq m \leq M)$ is less than $(r+1)^2$, then the optimal generalized power point is fair for user m. We have found a condition (based on the given channel capacities) which is helpful in determining the set of users for which the optimal generalized power point is fair, but which unfortunately does not fully characterize fairness as we shall see in section 6.1.7.1.

We also observe the influence of the parameter r in the above theorems concerning fairness. For large values of r, throughput is deemed more important by the analyst in choosing the function P_r. Since the optimal solution is fair if $C_1/C_M < (r+1)^2$, the set of capacities which yield a fair solution also becomes larger (the term $(r+1)^2$ increases when r does). Thus we are willing to give throughput to slower channels as r increases, because delay is not accorded as much importance.

6.1.7.1 Fairness Characterization Counterexample

For the case of two channels, Theorem 6.4 shows that the optimal solution when generalized power P_r is maximized is fair if and only if $C_1/C_2 < (r+1)^2$. Thus Theorem 6.4 gives a characterization of fairness for $M = 2$ in terms of the given channel capacities. For the case of M channels, Theorem 6.5 shows that the optimal solution is fair if the condition $C_1/C_M < (r+1)^2$ on the channel capacities holds. We now show that this property does not characterize fairness for this general case. That is, if the condition $C_1/C_M \geq (r+1)^2$ holds, then it is not necessarily the case that the optimal solution is unfair. Specifically, we prove the following

Theorem 6.9

For any real number $\alpha \geq (1+r)^2$, there is an M/M/1 parallel network with $C_1/C_M = \alpha$, but whose optimal solution for generalized power P_r is fair.

In fact, the example network we choose is of the form $C_1/C_i = \alpha$ for $2 \leq i \leq M$. For notational convenience, define $\beta \triangleq \sqrt{\alpha}$ so that $\beta^2 = \alpha$. Therefore, we have $\beta \geq 1+r$. We consider a parallel network of M channels with $C_1 = \alpha C = \beta^2 C$ and $C_2 = \cdots = C_M = C$. Thus the ratio of the capacity of the single fast channel to the capacity of any of the slow channels is $\alpha = \beta^2$. We have $S_1 = 1$ and $S_2 = \cdots = S_M = 1/\beta$. We will show that there is a value of $M = M'$ large enough such that the resulting parallel system yields a fair global solution point when generalized power P_r is maximized. Intuitively, this comes about since the accumulated throughput for a large number of slow channels is great enough to overcome the additional delay introduced by those slow channels.

To prove this theorem, we will show that

(i) v_M is real for large M

(ii) v_M yields a feasible interior solution point for large M

(iii) $\lim_{M \to \infty} P_r^M(v_M) = \infty$

Assuming (i), (ii) and (iii) hold, we now prove Theorem 6.9. We can find an integer M' such that, using (i) and (ii), $v_{M'}$ is real and yields a feasible interior solution point, and also, using (iii), such that

$$P_r^{M'}(v_{M'}) > P_r^j(v_j)$$

for all $j < M'$. Thus, as $P_r^j(v_j) = P_r^{M'}(v_j)$, we also have

$$P_r^{M'}(v_{M'}) > P_r^{M'}(v_j)$$

for $j < M'$. The optimal boundary point is given by v_j for some $j < M'$ (recall, the subnetworks of interest are obtained by dropping the slowest channel in a recursive manner), and thus the above choice of M' shows that this critical point yields higher generalized power than the optimal boundary point. Therefore, the solution corresponding to $v_{M'}$ must be globally optimal. From (i) and (ii), the root $v_{M'}$ is real and yields a feasible solution point which is an interior critical point of $P_r = P_r^{M'}$. Since this point is interior to the feasible region, it yields a *fair* solution, i.e., each user has non-zero throughput. This proves Theorem 6.9.

Before proving (i), (ii) and (iii), we establish some results which will assist us in these proofs. We first set $L \triangleq M - 1$, so that the network has 1 fast channel of capacity αC and L slow channels each of capacity C. We observe that $\sum_{i=1}^{M} S_i = 1 + \frac{L}{\beta}$ and $\sum_{i=1}^{M} S_i^2 = 1 + \frac{L}{\beta^2}$. The roots $u = u_M = u_{L+1}$ and $v = v_M = v_{L+1}$ are given in terms of the discriminant

$$D(L) = (r+2)^2 \left(\sum_{i=1}^{M} S_i\right)^2 - 4(r+1)(1+L)\sum_{i=1}^{M} S_i^2 \tag{6.51}$$

Using the values for S_i, we have

$$D(L) = (r+2)^2\left(1 + \frac{L}{\beta}\right)^2 - 4(r+1)(1+L)\left(1 + \frac{L}{\beta^2}\right)$$

or

$$D(L) = (r^2+4r+4)\left(1 + \frac{2L}{\beta} + \frac{L^2}{\beta^2}\right) - (4r+4)\left(1 + L + \frac{L}{\beta^2} + \frac{L^2}{\beta^2}\right)$$

Multiplying and collecting terms in powers of L, we obtain

$$D(L) = r^2 - \left[(4r+4) - (2r^2+8r+8)\frac{1}{\beta} + (4r+4)\frac{1}{\beta^2}\right]L + r^2\frac{L^2}{\beta^2}$$

which finally yields

$$D(L) = \frac{1}{\beta^2}\left[r^2L^2 - \left[(4r+4)\beta^2 - (2r^2+8r+8)\beta + (4r+4)\right]L + r^2\beta^2\right]$$

We may rewrite this expression in the form

$$D(L) = \frac{1}{\beta^2}\left[r^2L^2 - f(\beta)L + r^2\beta^2\right] \tag{6.52}$$

where we have defined

$$f(\beta) \triangleq (4r+4)\beta^2 - (2r^2+8r+8)\beta + (4r+4) \tag{6.53}$$

By differentiating equation (6.53) with respect to β, we observe that $f(\beta)$ is strictly convex with a global minimum at

$$\beta_{\min} = \frac{2r^2+8r+8}{8r+8} = 1 + \frac{2r^2}{8r+8}$$

Note that

$$\beta_{\min} = 1 + r\,\frac{r}{4(r+1)} < 1 + r$$

Therefore, $f(\beta)$ is a strictly convex *increasing* function of β for $\beta \geq 1+r$. Evaluating $f(\beta)$ at the point $\beta = 1+r$ yields

$$f(1+r) = 4(1+r)^3 - (2r^2+8r+8)(1+r) + 4(1+r)$$

or

$$f(1+r) = (1+r)\left[4(1+2r+r^2) - (2r^2+8r+8) + 4\right]$$

Therefore

$$f(1+r) = 2r^2(1+r) > 0$$

Since $f(\beta)$ is strictly increasing for $\beta \geq 1+r$, we must have

$$f(\beta) \geq f(1+r) = 2r^2(1+r) > 0$$

in this range. Thus $f(\beta) \geq 0$ for $\beta \geq 1+r$, and so $f(\beta)$ is positive for the given (fixed) value of β of Theorem 6.8.

Writing equations (6.28) and (6.29) in terms of the parameters of the example network(s) we are studying, we have

$$u = \frac{(r+2)(1+\frac{L}{\beta}) + \sqrt{D(L)}}{2(r+1)(1+L)} \tag{6.54}$$

and

$$v = \frac{(r+2)(1+\frac{L}{\beta}) - \sqrt{D(L)}}{2(r+1)(1+L)} \tag{6.55}$$

We know that the solution point corresponding to v gives higher generalized power, and thus we wish to evaluate $P_r(v)$ as given by equation (6.34). We first calculate

$$(r+1)u - v = \frac{(r+1)(r+2)(1+\frac{L}{\beta}) + (r+1)\sqrt{D(L)}}{2(r+1)(1+L)} - \frac{(r+2)(1+\frac{L}{\beta}) - \sqrt{D(L)}}{2(r+1)(1+L)}$$

or

$$(r+1)u - v = \frac{r(r+2)(\beta+L) + (r+2)\beta\sqrt{D(L)}}{2(r+1)(1+L)\beta}$$

Therefore

$$[(r+1)u - v]^r = \frac{(r+2)^r[r(\beta+L) + \beta\sqrt{D(L)}]^r}{2^r(r+1)^r(1+L)^r\beta^r}$$

and so, using equation (6.34), we have (recall that $M = 1+L$)

$$P_r(v) = (\mu C_1)^{r+1}\frac{r+1}{2^r\beta^r}v^{r+2}[r(\beta+L) + \sqrt{D(L)}]^r$$

Since $C_1 = \beta^2 C$, we obtain

$$P_r(v) = (\mu C)^{r+1}\frac{(r+1)\beta^{r+2}}{2^r}v^{r+2}[r(\beta+L) + \beta\sqrt{D(L)}]^r \tag{6.56}$$

We now prove (i). From equation (6.30), we see that the roots u and v are real if and only if $D(L) \geq 0$. From equation (6.52), we observe that $D(L)$ is positive for large L, and thus (i) holds.

We next prove (ii). We wish to show that, for large L, the root $v = v_M = v_{L+1}$ yields a feasible solution which is also an interior point of the feasible region. That is, we claim that $v < S_M$ for large L. This is true if and only if

$$v < \frac{1}{\beta} \tag{6.57}$$

for large L, which we now proceed to prove. Recall, from the proof of (i), that the roots u and v are real for large L. Using equation (6.55), we see that equation (6.57) is equivalent to

$$\frac{(r+2)\left(1+\frac{L}{\beta}\right)-\sqrt{D(L)}}{2(r+1)(1+L)} < \frac{1}{\beta}$$

or

$$(r+2)(\beta+L)-\beta\sqrt{D(L)} < 2(r+1)(1+L)$$

This last inequality yields

$$(r+2)\beta+(r+2)L < 2(r+1)+2(r+1)L+\beta\sqrt{D(L)}$$

or

$$(r+2)\beta < rL+2(r+1)+\beta\sqrt{D(L)}$$

which is clearly true for large L, since $D(L)$ is positive for large L, and β is fixed. This proves equation (6.57) and therefore (ii). Thus, for large L, we have shown that the root v is real and yields a feasible solution point which is an interior point of the feasible region. Therefore, it yields a *fair* solution.

We now prove (iii). We first find a lower bound on $v = v_{L+1} = v_M$ which is useful in bounding the corresponding value of generalized power $P_r(v)$. We claim that

$$\frac{1}{(r+1)\beta} < v \tag{6.58}$$

for large L. Recall that, for large L, v is real and yields a feasible solution which is fair. Using equation (6.55), we see that equation (6.58) is true if and only if

$$\frac{1}{(r+1)\beta} < \frac{(r+2)\left(1+\frac{L}{\beta}\right)-\sqrt{D(L)}}{2(r+1)(1+L)}$$

or

$$2(r+1)(1+L) < (r+1)(r+2)(\beta+L)-(r+1)\beta\sqrt{D(L)}$$

This is equivalent to

$$2(1+L) < (r+2)\beta+(r+2)L-\beta\sqrt{D(L)}$$

or

$$\beta\sqrt{D(L)} < rL+(r+2)\beta-2 \tag{6.59}$$

Note that, since $\beta \geq 1+r$, we have

$$(r+2)\beta \geq (r+2)(r+1) > 2$$

and so $(r+2)\beta-2 > 0$. Thus for equation (6.59) to hold, we need only show that

$$\beta\sqrt{D(L)} \leq rL \tag{6.60}$$

for large L. Since $D(L)$ is positive for large L, we may square equation (6.60) and preserve the inequality to obtain

$$\beta^2 D(L) \leq r^2 L^2$$

Using equation (6.52), we have

$$r^2L^2 - f(\beta)L + r^2\beta^2 \leq r^2L^2$$

Rearranging terms, we need only prove

$$r^2\beta^2 \leq f(\beta)L$$

for large L. Since $f(\beta)$ is positive for $\beta \geq 1+r$, this inequality clearly holds for large L (recall that β is fixed). This proves equation (6.58).

From the above results, for large L (i.e., for large M), the root $v = v_{L+1} = v_M$ is real, yields a feasible solution point which is fair, and also satisfies equation (6.58). In particular, we have shown for large L that

$$\frac{1}{(r+1)\beta} < v < \frac{1}{\beta} \tag{6.61}$$

We now bound the value $P_r(v)$ for large L. Recall that equation (6.56) gives $P_r(v)$ as

$$P_r(v) = (\mu C)^{r+1} \frac{(r+1)\beta^{r+2}}{2^r} v^{r+2} \left[r(\beta+L) + \beta\sqrt{D(L)}\right]^r$$

We now use the bound given in equation (6.58) to obtain

$$P_r(v) > (\mu C)^{r+1} \frac{(r+1)\beta^{r+2}}{2^r} \cdot \frac{1}{[(r+1)\beta]^{r+2}} \left[r(\beta+L) + \beta\sqrt{D(L)}\right]^r$$

or

$$P_r(v) > (\mu C)^{r+1} \frac{\left[r(\beta+L) + \beta\sqrt{D(L)}\right]^r}{2^r(r+1)^{r+1}} \tag{6.62}$$

for large L. Since $D(L)$ is positive for large L, the right-hand side of the above inequality increases without bound as $L \to \infty$ (i.e., as $M \to \infty$). Therefore, we have

$$\lim_{M \to \infty} P_r^M(v_M) = \infty \tag{6.63}$$

where P_r^M is the generalized power function of an M channel parallel network. This proves (iii), and thus Theorem 3.9.

This concludes our analysis of the generalized power function which extends Giessler's power for the M/M/1 parallel network. We have found a solution procedure for the general case of M channels, and we have derived the exact analytical solution for the two channel net. The solution of this optimization problem was not an easy task, however, and several negative aspects of the optimal operating point were revealed. Simple two channel examples of parallel nets with an unfair optimal operating point or such that the local generalized power point differs from the global generalized power point were presented. In chapter 5, it was also shown that, for $r > 1$, generalized power is not concave for the M/M/1 queueing system (the case of one channel). The difficulty of the method of solution and the undesirable properties which resulted suggest that the generalized power problem for general computer network configurations may perhaps be hard to solve and yield poor operating points for this particular definition. However, a generalized power function extending one first introduced by Kleinrock turns out to not possess many of these undesirable qualities.

6.2 Generalised Power (Extension of P_K)

We now choose to analyze the M/M/1 parallel network (with unknown routing) by using a family of generalized power functions based on those first introduced by Kleinrock [Klei79] for a single queueing system. His definition is

$$P_{K,r} = \frac{\rho^r}{T/\bar{x}} \tag{6.64}$$

In this section, the notation $P_{K,r}$ will refer to the function defined in equation (6.64), while the notation P_r will continue to refer to the extension of Giessler's power function defined as $P_{G,r}$ in equation (6.1) and analyzed in the first part of this chapter. The case $r = 1$ of $P_{K,r}$ was extended to a definition valid for an arbitrary network in chapter 4. Proceeding in the same way, we may extend equation (6.64) to a general network with M channels as

$$P_{K,r} = (\gamma \bar{x})^r \frac{\bar{x}}{T} = \frac{(\gamma \bar{x})^{r+1}}{\gamma T}$$

This becomes

$$P_{K,r} = \frac{\left(\sum_{i=1}^{M} \lambda_i \bar{x}_i\right)^{r+1}}{\sum_{i=1}^{M} \lambda_i T_i} \tag{6.65}$$

or

$$P_{K,r} = \frac{\left(\sum_{i=1}^{M} \rho_i\right)^{r+1}}{\sum_{i=1}^{M} \bar{N}_i} \tag{6.66}$$

In a single server queueing system, ρ (the fraction of time the server is busy), is also equal to the expected number of busy servers. In the general network, this latter expectation is simply $\sum_{i=1}^{M}\rho_i$. We have also replaced $\bar{x}$ with $\sum_{i=1}^{M}\frac{\lambda_i}{\gamma}\bar{x}_i$ as in chapter 4. Since $\bar{N}_i=\rho_i/(1-\rho_i)$ for M/M/1, equation (6.66) may also be written solely in terms of ρ_i as

$$P_{K,r}=\frac{\left(\sum_{i=1}^{M}\rho_i\right)^{r+1}}{\sum_{i=1}^{M}\frac{\rho_i}{1-\rho_i}} \tag{6.67}$$

We now solve the parallel network problem of chapter 3 using the objective function $P_{K,r}$. Recall that the optimal solution of this network for P_r (determined in the previous part of this chapter) was unfair, was quite complicated to derive, and had the property that local generalized power was not equal to global generalized power. The model we consider is that of an M/M/1 parallel network with M channels and no restriction on the routing of messages through the net. We wish to find channel flows λ_i for $1\le i\le M$ which maximize $P_{K,r}$. This yields a multi variable optimization problem with M unknowns. We will find all critical points of $P_{K,r}$ (points with $\nabla P_{K,r}=0$) and compare them with the maximal boundary point as in chapter 4. To this end, we first find the M partial derivatives of $P_{K,r}$ with respect to the channel flows. Taking the partial derivative of $P_{K,r}$ with respect to λ_j from equation (6.65) gives

$$\frac{\partial P_{K,r}}{\partial\lambda_j}=\frac{(r+1)\left(\sum_{i=1}^{M}\lambda_i T_i\right)\left(\sum_{i=1}^{M}\lambda_i\bar{x}_i\right)^{r}\bar{x}_j-\left(\sum_{i=1}^{M}\lambda_i\bar{x}_i\right)^{r+1}\left(\lambda_j\frac{dT_j}{d\lambda_j}+T_j\right)}{\left(\sum_{i=1}^{M}\lambda_i T_i\right)^{2}}$$

or

$$\frac{\partial P_{K,r}}{\partial\lambda_j}=\frac{\left(\sum_{i=1}^{M}\lambda_i\bar{x}_i\right)^{r+1}}{\left(\sum_{i=1}^{M}\lambda_i T_i\right)^{2}}\left[\frac{(r+1)\sum_{i=1}^{M}\lambda_i T_i}{\sum_{i=1}^{M}\lambda_i\bar{x}_i}\bar{x}_j-\left(\lambda_j\frac{dT_j}{d\lambda_j}+T_j\right)\right]$$

If we assume that $\nabla P_{K,r}=0$, then $\partial P_{K,r}/\partial\lambda_j=0$ for all $1\le j\le M$. Thus we have (using $\bar{x}_j=1/\mu C_j$)

$$(\mu C_j)\left(\lambda_j\frac{dT_j}{d\lambda_j}+T_j\right)=\frac{(r+1)\sum_{i=1}^{M}\lambda_i T_i}{\sum_{i=1}^{M}\lambda_i\bar{x}_i}=(\mu C_k)\left(\lambda_k\frac{dT_k}{d\lambda_k}+T_k\right)\qquad\forall\, j,k \tag{6.68}$$

Since we assume that each channel acts as an M/M/1 queueing system, we may use equation (6.6) which states that

$$\lambda_j \frac{dT_j}{d\lambda_j} + T_j = \frac{1}{\mu C_j} \cdot \frac{1}{(1-\rho_j)^2}$$

Substituting this expression into equation (6.68) yields

$$\frac{1}{(1-\rho_j)^2} = (r+1) \frac{\sum_{i=1}^{M} \lambda_i T_i}{\sum_{i=1}^{M} \lambda_i \bar{x}_i} = \frac{1}{(1-\rho_k)^2} \qquad \forall j, k \tag{6.69}$$

Therefore, for $\nabla P_{K,r} = 0$, we have

$$\rho_j = \rho_k \qquad \forall j, k \tag{6.70}$$

and thus

$$\bar{N}_j = \bar{N}_k \qquad \forall j, k \tag{6.71}$$

Using equations (6.70) and (6.71), we find that equation (6.69) simplifies to

$$(r+1) \frac{M \bar{N}_1}{M \rho_1} = \frac{1}{(1-\rho_1)^2}$$

Since $\bar{N} = \rho/(1-\rho)$ for M/M/1, we find

$$\frac{r+1}{1-\rho_1} = \frac{1}{(1-\rho_1)^2}$$

or

$$\rho_1 = \frac{r}{r+1}$$

and

$$\bar{N}_1 = r$$

Thus $\rho_i = r/(r+1)$ and $\bar{N}_i = r$ for all $1 \le i \le M$ at this (unique) critical point of the function $P_{K,r}$, and the corresponding function value is

$$P_{K,r} = \frac{[Mr/(r+1)]^{r+1}}{Mr} = \frac{(Mr)^r}{(r+1)^{r+1}}$$

As in chapter 4, the optimal boundary point occurs when one of the λ_i is zero. Therefore we consider a parallel network with $M-1$ channels, and by the above analysis, the optimal boundary point will be such that all the other $\rho_i = r/(r+1)$. Thus the optimal boundary point will yield a function value of

$$P_{K,r} = \frac{[(M-1)r]^r}{(r+1)^{r+1}}$$

and so the critical point of $P_{K,r}$ is globally optimal. We have shown the following

Theorem 6.10

For the M/M/1 parallel network (with unknown routing), the channel utilizations which maximize the generalized power function $P_{K,r}$ are

$$\rho_i^* = \frac{r}{r+1} \qquad 1 \leq i \leq M \tag{6.72}$$

We also have the following

Theorem 6.11

For the M/M/1 parallel network (with unknown routing), the average number of messages at each channel when $P_{K,r}$ is maximized is

$$\bar{N}_i^* = r \qquad 1 \leq i \leq M \tag{6.73}$$

This theorem extends Kleinrock's result for M/M/1 to the M/M/1 parallel system with unknown routing. We note that the undesirable properties of the optimal solution for this network when P_r was maximized have vanished. The optimal solution for $P_{K,r}$ is fair (it is as fair as possible), while global generalized power and local generalized power are now the same. The utilizations of the individual channels have taken on greater weight in the objective function. This is reflected in the optimal solution, because every channel is utilized at the same level regardless of the speed of the channel.

Let us now consider formulations PF1, PF2, and PF3 with objective function $P_{K,r}$ for a general network topology. The optimization problem for the parallel network considered above is an instance of PF1, and the only constraints of this problem are that each ρ_i be between zero and one. We note that the feasible region for that particular problem consists of the set of M dimensional vectors $\rho = (\rho_1, \ldots, \rho_M)$ such that $0 \leq \rho_i \leq 1$ for all $1 \leq i \leq M$, while the objective function is given solely in terms of the ρ_i in equation (6.67). Formulations PF1, PF2, and PF3 for general network topologies will have the same objective function as given in equation (6.67), but the feasible region will be a subset of the feasible region for the parallel network. The constraints imposed by the particular topology of the network and/or by the formulation being considered (given routing or relative traffic matrix) will reduce the set of ρ_i which are feasible. If we consider an arbitrary network and relax (disregard) these extra constraints, we obtain the parallel network optimization problem. Thus if the optimal solution to the parallel network is feasible for the particular network we are examining, then it must be optimal for this other network also. That is, if the vector given by $\rho_i = r/(r+1)$ for all i is *feasible* for a particular network problem, then it is *optimal*. This says that, if it is at all possible to have $\bar{N}_i = r$ based upon the constraints imposed by the topology and/or formulation, then the optimal policy is to do just that. As in chapter 4, we can give several examples of network problems such that $\bar{N}_i = r$ for all i is optimal. We have shown that the parallel network with arbitrary channel

capacities is such an example, while the series network with equal channel capacities is another obvious example. A third interesting example is a unidirectional ring with M nodes and M channels, all channels having the same capacity C. As in chapter 4, this third example has both a fair and an unfair optimal generalized power solution. Thus, even if we "keep the pipe full", the resulting optimal solution using Kleinrock's generalized power function $P_{K,r}$ may still be unfair.

In this section we have introduced a family of generalized power functions based upon those introduced by Kleinrock. We have seen that the optimization of these functions yields the rule of thumb that (to maximize $P_{K,r}$) one should always operate a network in order to have $\overline{N}_t = r$ as long as the constraints of the network topology and problem formulation allow this to be done. However, many networks do not permit such a "keep the pipe full" solution. A general solution procedure for such networks is thus far unknown and may be difficult to discover.

CHAPTER 7
Deterministic Rules of Thumb

In the previous chapters we analyzed several throughput-delay tradeoff functions (power and its relatives) for various computer network configurations and problem formulations. The models considered in those chapters were used to represent conventional wire networks of the ARPANET type, and consequently all queueing systems which were examined were single-server systems (the server being the message channel) with no blocking of messages (infinite buffer size). A natural extension of that analysis is to consider models which contain multiple-server queueing systems. In fact, so-called Markovian networks (both open and closed) which consist of multiple-server systems have been extensively analyzed and used to model various computer phenomena [Klei76]. One example of the use of such models is in the area of telephone systems [Sysk60], for which the multiple-server assumption is particularly applicable. A second area of application is that of the performance analysis of computer systems behavior [Saue81, Lave83]. Another extension of the previous work is to consider models which incorporate blocking (finite buffer size). Let us now examine multiple-server systems (with and without blocking) to find the "appropriate operating point" using power.

7.1 Pure Delay Systems

In attempting to extend Kleinrock's result that $\overline{N^*} = 1$ for M/G/1 to the multiple-server system M/G/m, we run up against a serious problem. The M/G/m system is extremely difficult to analyze. In fact, general expressions for system variables (such as $\overline{N}$, W, T, etc.) are unknown. Approximations can be made with various degrees of success, but an exact analysis has not been published. If we turn our attention to the M/M/m queueing system, much more is known. As mentioned above, networks of these systems have been studied, but we consider only a single M/M/m system in what follows. As we shall see, even this case is difficult to analyze in terms of power, because the closed-form expressions for various system parameters (including T) become quite complicated as the number of servers m gets large.

Let us consider, then, an M/M/m system. The arrival rate to this m-server system is Poisson with rate λ, and service time is exponentially distributed with mean $\overline{x} = 1/\mu$. Note that the throughput of this pure delay system is simply $\gamma = \lambda$. The average time in system, T, is given by [Klei76]

$$T = \overline{x} + \overline{x}\,\frac{Q_m}{m(1-\rho)} \tag{7.1}$$

where $\rho = (\gamma \bar{x})/m$ is the expected fraction of busy servers, while Q_m is the probability that all m servers are busy (and thus Q_m is the probability that the system contains m or more). Note that Q_m is also the probability that an arrival to the system must queue, since we have Poisson arrivals. Further, Q_m is given by the (complicated) expression

$$Q_m = \frac{\frac{(m\rho)^m}{m!(1-\rho)}}{\sum_{k=0}^{m-1} \frac{(m\rho)^k}{k!} + \frac{(m\rho)^m}{m!(1-\rho)}} \tag{7.2}$$

We note, for example, that $Q_1 = \rho$ and $Q_2 = (2\rho^2)/(1+\rho)$. Using Little's result, we see from equation (7.1) that the average number in system satisfies

$$\bar{N} = m\rho + \frac{\rho}{1-\rho} Q_m \tag{7.3}$$

Since $\bar{x}$ is constant for this system, one might be tempted to conclude that all the power functions defined in chapter 4 above (P_G, P_N, and P_K) will yield the same optimal power point. We now show that this is indeed the case. The power function introduced by Kleinrock [Klei79] is

$$P_K = \frac{\rho}{T/\bar{x}} = \rho \frac{\bar{x}}{T} = \frac{\gamma}{mT}(\bar{x})^2 = P_G \frac{(\bar{x})^2}{m}$$

where

$$P_G = \frac{\gamma}{T}$$

is Giessler's power function. For multiple server systems, note that Kleinrock's function P_K depends on the number of servers m through the term $\rho = (\gamma \bar{x})/m$. But, since m is constant, P_K will be optimized at the same value γ^* for which P_G (and also P_N) is optimized.

Arguing as in chapter 2 for the case of a single-server queueing system, we see that the optimal value of any of the above notions of power can be determined by solving the familiar equation

$$T^* = \gamma^* \left. \frac{dT}{d\gamma} \right|_{\gamma=\gamma^*} \tag{7.4}$$

or, equivalently,

$$\bar{N}^* = (\gamma^*)^2 \left. \frac{dT}{d\gamma} \right|_{\gamma=\gamma^*} \tag{7.5}$$

In order to use equations (7.4) or (7.5), we must first find Q_m (and then T). For $m = 1$, since $Q_1 = \rho$, equation (7.1) yields $T = \bar{x}/(1-\rho)$. If we solve equation (7.5) for $m = 1$ using this value for T, we find that $\bar{N}^* = 1$ and $\rho^* = 1/2$, which agrees with Kleinrock's results for M/M/1.

Let us now find the optimal power point for the case of two servers ($m = 2$). Using the value $Q_2 = (2\rho^2)/(1+\rho)$ given above, equation (7.1) yields

$$T = \frac{\bar{x}}{1-\rho^2} \tag{7.6}$$

From this expression for T and Little's result (recall that $\rho = (\gamma\bar{x})/2$ for M/M/2), we also have

$$\bar{N} = \frac{\gamma\bar{x}}{1-\rho^2} = \frac{2\rho}{1-\rho^2} \tag{7.7}$$

Differentiating equation (7.6), we obtain

$$\frac{dT}{d\gamma} = \bar{x}\,\frac{2\rho(\bar{x}/2)}{(1-\rho^2)^2} = \rho T^2$$

Using equation (7.5), we see that

$$\bar{N}^* = (\gamma^*)^2\rho^*(T^*)^2 = \rho^*(\bar{N}^*)^2$$

Therefore

$$\rho^*\bar{N}^* = 1 \tag{7.8}$$

for the M/M/2 system at maximal power. Equations (7.7) and (7.8) now yield

$$\frac{2(\rho^*)^2}{1-(\rho^*)^2} = 1$$

or

$$3(\rho^*)^2 = 1$$

Solving this equation, we find

$$\rho^* = \frac{\sqrt{3}}{3} \tag{7.9}$$

and

$$\bar{N}^* = \sqrt{3} \tag{7.10}$$

for the system M/M/2. This yields a counterexample to the tantalizing conjecture that $\bar{N}^* = m$ for M/M/m. Also note that the optimal value of Kleinrock's power function is

$$P_K^* = \frac{\rho^*}{T^*/\bar{x}} = \rho^*[1-(\rho^*)^2]$$

or

$$P_K^* = \frac{\sqrt{3}}{3}\cdot\frac{2}{3} = \frac{2\sqrt{3}}{9}$$

For $m > 2$, Q_m becomes quite complicated, and thus T and the power function do also. Values of system variables have been numerically calculated for various values of m by Kleinrock, and plots of the results appear in [Klei79]. A plot of power P_K versus ρ for different values of m is given in Figure 3.5 of that paper, and it shows that

$$\lim_{m \to \infty} \rho^* = 1 = \lim_{m \to \infty} P_K^*$$

for M/M/m. A plot of $\overline{N}^*$ versus m in Figure 3.6 of the same paper indicates that

$$\frac{\overline{N}^*}{m} \leq 1$$

for all m, and that

$$\lim_{m \to \infty} \frac{\overline{N}^*}{m} = 1$$

However, the above results for ρ^* and $\overline{N}^*$ have so far resisted an analytical proof.

Another plot in that paper (Figure 3.4) gives the normalized average time in system $T/\overline{x}$ versus the efficiency ρ for different values of m. This plot indicates that, in the limit, T behaves exactly as in the deterministic D/D/1 system. This turns out to be only one of a number of cases of deterministic behavior that occur due to the smoothing effect of the law of large numbers which are considered in [Klei79]. This smoothing principle for large shared systems is as follows.

A large population presents a total demand which is deterministic with a value equal to the sum of the average demands of each member of the population (as opposed to the sum of the peak demands of each).

Kleinrock studies this principle (and its interaction with power) in relation to several queueing system models. In the next two sections of this chapter we review these results and provide new proofs of theorems which appear in [Klei79]. Moreover, we correct the value obtained at a particular point in several equations of this paper. The results of the next two sections illustrate the deterministic behavior of large systems, and we apply this analysis to a model of a packetized voice network in the final section of this chapter.

7.2 Pure Loss Systems

Consider m data channels which are fed by a Poisson input source of messages with rate λ. Assume that any message which arrives to find all m channels busy is lost (blocked). We model this as an M/G/m/m pure loss system (see Figure 7.1).

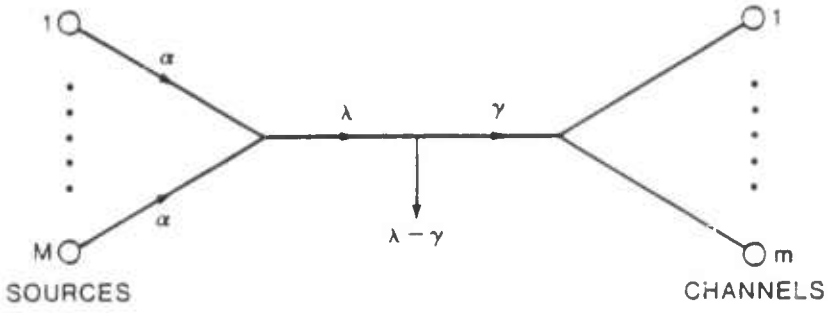

Figure 7.1 The M/G/m/m System

Unlike the M/G/m system, expressions for variables of the M/G/m/m system are known. As might be expected, however, these expressions become complicated for large values of m. But in the limit as $m \to \infty$, the law of large numbers takes hold and deterministic behavior results. To study this behavior, let us decompose our Poisson source into M sources, each generating messages at a Poisson rate α. We thus have $\lambda = M\alpha$. The service time distribution is assumed to be general with mean $\bar{x}$. As before, $\bar{x} = \bar{b}/C$ where $\bar{b}$ is the mean message length in bits and C is the channel capacity in bits per second. Further define $a \triangleq \alpha\bar{x}$ and $A \triangleq \lambda\bar{x} = M\alpha\bar{x} = Ma$. In the language of telephony theory, a is the load offered by each of the M sources, while A is the total offered load. We also define the offered load per channel δ as

$$\delta \triangleq \frac{A}{m} = \frac{\lambda\bar{x}}{m} = \frac{M}{m}\alpha\bar{x} = \frac{M}{m}a$$

Thus δ is simply a multiple of M/m, the ratio of sources to channels.

Performance measures of interest are given in terms of the blocking probability $B = B_m(A)$. For example, the arrival rate of messages which actually enter the system (which is identical with the throughput γ) is simply $\lambda(1-B)$. Thus the arrival rate λ and the throughput γ are not the same for this loss system. The total carried load is therefore $\gamma\bar{x} = \lambda(1-B)\bar{x} = A(1-B)$, and the efficiency ρ of each channel is

$$\rho = \frac{\gamma\bar{x}}{m} = \frac{\lambda(1-B)\bar{x}}{m} = \frac{A(1-B)}{m}$$

Note that

$$\rho = \delta(1-B)$$

Since $T = \bar{x}$, we see that the variable of interest is B, the blocking probability. Once we know B, other parameters (such as throughput and power) may be calculated. The blocking probability B for the M/G/m/m system is given by Erlang's famous formula

$$B = B_m(A) = \frac{A^m/m!}{\sum_{j=0}^{m} A^j/j!} \tag{7.11}$$

This is easily proved for M/M/m/m; for a rigorous proof of the general service time case see the book by Cohen [Cohe76]. If $m = 1$ (so that $A = \delta$), we have $B = \delta/(1+\delta)$. However, the formula for B becomes quite complicated for large m. But a simple limiting behavior is exhibited in [Klei79] for fixed values of M/m. This key variable, the ratio of sources to channels, appears in the limiting expression. Specifically, we keep the variable δ constant (and thus the ratio M/m constant) as $m \to \infty$ and examine the behavior of B.

We first give a new proof to the following theorem from [Klei79], which characterizes the limiting behavior of the blocking probability as both the number of users (sources) and the number of servers (channels) increase in the same proportion.

Theorem 7.1 (Kleinrock)

For the M/G/m/m system, as δ remains constant and $m \to \infty$ we have

$$\lim_{m \to \infty} B = \begin{cases} 0 & 0 \le \delta \le 1 \\ 1 - \dfrac{1}{\delta} & 1 \le \delta \end{cases} \tag{7.12}$$

Before proving this result, we remark that, for $t \ge 0$,

$$\sum_{j=0}^{m} \frac{t^j e^{-t}}{j!} = \int_t^{\infty} \frac{x^m e^{-x}}{m!}\, dx \tag{7.13}$$

which may be shown by induction on m and integration by parts. Equation (7.13) may also be easily justified in the following manner. Consider a random variable $\hat{x}$ which has an Erlang-$(m+1)$ distribution with parameter η. That is, $\hat{x}$ is the sum of $m+1$ independent random variables each of which is exponentially distributed with mean $1/\eta$. The density $f(x)$ of $\hat{x}$ is

$$f(x) = \frac{\eta(\eta x)^m e^{-\eta x}}{m!} \qquad x \ge 0$$

and the distribution function $F(x)$ is

$$F(x) = 1 - \sum_{j=0}^{m} \frac{(\eta x)^j e^{-\eta x}}{j!} \qquad x \ge 0$$

Setting $\eta = 1$, we observe that equation (7.13) holds by noting that both sides represent the probability $1 - F(t) = P[\hat{x} > t] = \int_t^{\infty} f(x)\, dx$.

Let us now prove Theorem 7.1. It is easier to consider $1/B$. We write

$$\frac{1}{B} = \frac{\sum_{j=0}^{m} A^j / j!}{A^m / m!} = \frac{\sum_{j=0}^{m} A^j e^{-A} / j!}{A^m e^{-A} / m!}$$

where we obtained the right-hand term by multiplying the numerator and the denominator of the middle expression by e^{-A}. Using equation (7.13), we have

$$\frac{1}{B} = \frac{\int_A^\infty \frac{x^m e^{-x}}{m!} dx}{A^m e^{-A} / m!} = \int_A^\infty \left(\frac{x}{A}\right)^m e^{-(x-A)} dx$$

Making the change of variable $y = x - A$ yields

$$\frac{1}{B} = \int_0^\infty \left(1 + \frac{y}{A}\right)^m e^{-y} dy$$

We now wish to determine the limiting behavior of the above expression as $m \to \infty$. Writing $A = m\delta$ gives

$$\frac{1}{B} = \int_0^\infty \left(1 + \frac{y/\delta}{m}\right)^m e^{-y} dy$$

We now wish to determine the limiting behavior of the above expression as $m \to \infty$. We first define the functions (for $m = 1, 2, \ldots$)

$$f_m(y) \triangleq \left(1 + \frac{y/\delta}{m}\right)^m e^{-y}$$

and also define

$$f(y) \triangleq e^{y/\delta} e^{-y} = e^{-y\left(1 - \frac{1}{\delta}\right)}$$

We observe that the sequence of functions $f_m(y)$ increases monotonically for all $0 \le y < \infty$ and converges to the function $f(y)$. Thus we may use the Monotone Convergence Theorem to interchange limit and integral, so that

$$\lim_{m \to \infty} \frac{1}{B} = \lim_{m \to \infty} \int_0^\infty f_m(y) dy = \int_0^\infty \lim_{m \to \infty} f_m(y) dy$$

$$= \int_0^\infty f(y) dy = \int_0^\infty e^{-y\left(1 - \frac{1}{\delta}\right)} dy$$

Thus

$$\lim_{m \to \infty} \frac{1}{B} = \begin{cases} \infty & 0 \le \delta \le 1 \\ \dfrac{1}{1 - \frac{1}{\delta}} & 1 < \delta \end{cases}$$

Therefore we have

$$\lim_{m \to \infty} B = \begin{cases} 0 & 0 \leq \delta \leq 1 \\ 1 - \dfrac{1}{\delta} & 1 \leq \delta \end{cases}$$

which proves Theorem 7.1.

From this result we see that the probability of success $P_S \triangleq 1 - B$ satisfies

$$\lim_{m \to \infty} P_S = \begin{cases} 1 & 0 \leq \delta \leq 1 \\ \dfrac{1}{\delta} & 1 \leq \delta \end{cases} \tag{7.14}$$

Note the deterministic behavior in the limiting case of m. As long as the offered load per channel can be handled by the system $(0 \leq \delta \leq 1)$, there is no blocking and all messages are transmitted. However, if the average amount of traffic is greater than the channels can handle $(1 < \delta)$, only a fraction $1/\delta$ of the messages are successful. The success probability P_S has been plotted versus δ for various values of m in Figure 7.2. We note that the curve for $m = 100$ is already very close to the limiting behavior exhibited in Theorem 7.1.

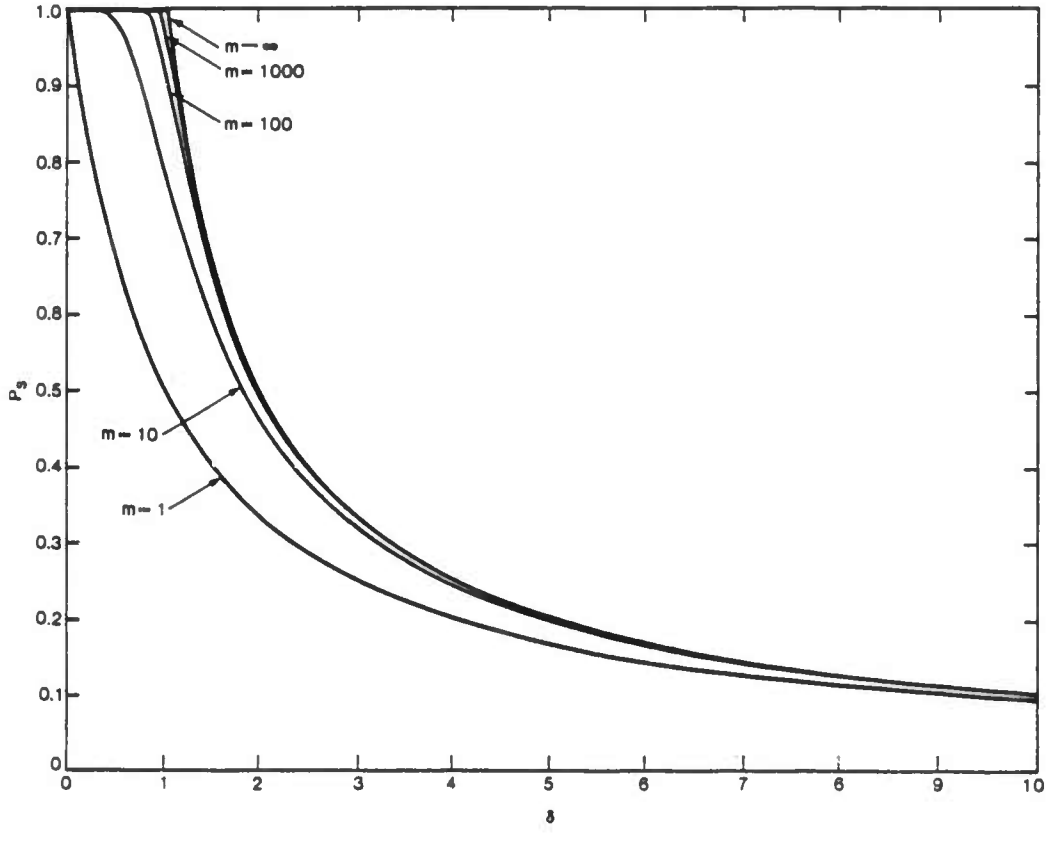

Figure 7.2 Success Probability for Various Values of m

As commented by Kleinrock [Klei79], a closer examination of the theorem indicates that the limiting blocking probability is zero for $0 \leq \delta \leq 1$ and then behaves like that of a system with $m = 1$ (recall that $B = \delta/(1+\delta)$ for M/G/1/1). Thus it appears as if the limiting pure loss system (for $1 \leq \delta$) acts like the M/G/1/1 system with a load of $\delta - 1$ with regard to blocking probability. To analyze this limiting behavior further, let us consider the probability p_{m-1} of there being $m-1$ in the system (recall that $B = p_m$). Then we have

$$p_{m-1} = \frac{A^{m-1}/(m-1)!}{\sum_{j=0}^{m} A^j/j!} = \frac{m}{A} p_m = \frac{1}{\delta} p_m$$

Thus as δ remains constant and $m \to \infty$

$$\lim_{m \to \infty} p_{m-1} = \begin{cases} 0 & 0 \leq \delta \leq 1 \\ \frac{1}{\delta} - \frac{1}{\delta^2} & 1 \leq \delta \end{cases}$$

and

$$\lim_{m \to \infty} [p_m + p_{m-1}] = \begin{cases} 0 & 0 \leq \delta \leq 1 \\ 1 - \frac{1}{\delta^2} & 1 \leq \delta \end{cases}$$

Similarly we may fix k and consider p_{m-k}. Then

$$p_{m-k} = \frac{A^{m-k}/(m-k)!}{\sum_{j=0}^{m} A^j/j!} = \frac{m(m-1)\cdots(m-k+1)}{A^k} p_m$$

or (as $A = m\delta$)

$$p_{m-k} = \left(\frac{1}{\delta}\right)^k \left(1 - \frac{1}{m}\right) \cdots \left(1 - \frac{k-1}{m}\right) p_m$$

Therefore

$$\lim_{m \to \infty} p_{m-k} = \begin{cases} 0 & 0 \leq \delta \leq 1 \\ \frac{1}{\delta^k} - \frac{1}{\delta^{k+1}} & 1 \leq \delta \end{cases}$$

and

$$\lim_{m \to \infty} [p_m + p_{m-1} + \cdots + p_{m-k}] = \begin{cases} 0 & 0 \leq \delta \leq 1 \\ 1 - \frac{1}{\delta^k} & 1 \leq \delta \end{cases}$$

We note that, for $1 < \delta$, the system is almost always in those states in which most of the servers are busy. But although the blocking probability for the limiting system behaves as that

of the M/G/1/1 system with the load per server shifted over by one unit, we see that the overall characteristics of the two systems are different. In particular, for fixed k, there is positive probability $1/\delta^k$ that more than k servers are idle. Thus the $m \to \infty$ system does not act exactly as an M/G/1/1 system (except for the case when δ is large).

Limiting values of other system parameters may now be easily calculated. Using the fact that $\rho = \delta(1 - B)$, we obtain the following

Theorem 7.2 (Kleinrock)

For the M/G/m/m system, as δ remains constant and $m \to \infty$ we have

$$\lim_{m \to \infty} \rho = \begin{cases} \delta & 0 \leq \delta \leq 1 \\ 1 & 1 \leq \delta \end{cases} \tag{7.15}$$

We also note that $\overline{N} = \gamma T = \gamma \overline{x}$ for this pure loss system, and so

$$\frac{\overline{N}}{m} = \frac{\gamma \overline{x}}{m} = \rho$$

Thus we have the following

Theorem 7.3 (Kleinrock)

For the M/G/m/m system, as δ remains constant and $m \to \infty$ we have

$$\lim_{m \to \infty} \frac{\overline{N}}{m} = \begin{cases} \delta & 0 \leq \delta \leq 1 \\ 1 & 1 \leq \delta \end{cases} \tag{7.16}$$

Theorems 7.2 and 7.3 were first shown by Kleinrock [Klei79].

We now examine the behavior of the power functions P_K and P_G for these pure loss systems. Since the delay remains constant ($T = \overline{x}$), the power function P_K satisfies

$$P_K = \frac{\rho}{T/\overline{x}} = \rho = \frac{\gamma \overline{x}}{m}$$

for M/G/m/m. Similarly we have

$$P_G = \frac{\gamma}{T} = \frac{\gamma}{\overline{x}}$$

In either case, the value of power is simply a constant multiple of throughput. When maximizing power for these pure loss systems, our decision variable is the input rate λ (or equivalently the offered load per channel δ) which corresponds to the traffic matrix $\{\gamma_{jk}\}$ of chapter 4.

An examination of the M/G/1/1 system (the case $m = 1$) now reveals an unfortunate property of power for these blocking systems if we use the previous definitions. Although the blocking probability does appear in the calculation of throughput as $\gamma = \lambda(1 - B)$, we see below that γ (and thus power) will continue to increase as the input rate λ increases, even though more and more messages are blocked. For the M/G/1/1 system, $\delta = A = \lambda\bar{x}$, and so equation (7.11) yields $B = \delta/(1+\delta)$ and $1 - B = 1/(1+\delta)$. Thus

$$\gamma = \lambda(1 - B) = \frac{\delta}{\bar{x}}\left(\frac{1}{1+\delta}\right) = \frac{\delta}{1+\delta} \cdot \frac{1}{\bar{x}}$$

so that

$$\gamma = \frac{B}{\bar{x}}$$

Both the throughput γ and the blocking probability B are strictly increasing as the input rate λ (and thus the offered load $A = \delta$) is increased. For M/G/1/1, since $P_G = \gamma/\bar{x}$ and $P_K = \gamma\bar{x}$, they are maximized as the design variable $\delta \to \infty$. But such an operating point also maximizes the blocking probability B, an unacceptable situation. Thus these notions of power are increasing functions of the favorable measure throughput, but are not decreasing functions of the undesirable quantity B, as they should be.

To alleviate this problem, Kleinrock [Klei79] proposed a power measure which incorporates the blocking probability. His definition is (again we use the notation P_K)

$$P = P_K = \frac{\rho(1 - B)}{T/\bar{x}} \tag{7.17}$$

Note that this new power function P_K is an increasing function of the throughput γ, a decreasing function of the delay T, and also a decreasing function of the blocking probability B. For the case of an M/G/m/m pure loss system, since $T = \bar{x}$, this new definition of power simplifies to $P = \rho(1 - B)$. If $m = 1$ (M/G/1/1), we have

$$P = \rho(1 - B) = \gamma\bar{x}(1 - B) = B(1 - B) = \left(\frac{\delta}{1+\delta}\right)\left(\frac{1}{1+\delta}\right)$$

or

$$P = \frac{\delta}{(1+\delta)^2}$$

This equation expresses P solely in terms of the design variable δ. Differentiating P with respect to δ gives

$$\frac{dP}{d\delta} = \frac{(1+\delta)^2 - 2\delta(1+\delta)}{(1+\delta)^4} = \frac{1 - \delta}{(1+\delta)^3}$$

and so P is maximized when $\delta = 1$ for the M/G/1/1 system. Note that $P^* = 1/4$, which is the same as for the M/M/1 pure delay system when we use Kleinrock's power function.

It remains to calculate the limiting value of this power function when $m \to \infty$. The following theorem, first proved by Kleinrock [Klei79], shows that this new definition of power gives a function which peaks at the (intuitively) correct point.

Theorem 7.4 (Kleinrock)

For the M/G/m/m system, as δ remains constant and $m \to \infty$ we have

$$\lim_{m \to \infty} P = \begin{cases} \delta & 0 \leq \delta \leq 1 \\ \dfrac{1}{\delta} & 1 \leq \delta \end{cases} \tag{7.18}$$

The maximum occurs when $\delta = 1$, i.e., where the total offered load per channel is unity. For small values of δ, there is no blocking, and an increase in the input rate results in increased throughput and increased power. For large values of δ, a decrease in the input rate results in a decrease in the blocking probability and an increase in power (even though the throughput decreases). Thus we have found simple expressions for the limiting behavior of the system parameters of interest. These results will be extended when we model a packetized voice network later in this chapter.

We close this section with the observation that there does not seem to be a nice intuitive explanation of the optimum power point for these systems with blocking. That is, the analogue of the "knee of the curve" rule seems to be absent for pure loss systems. For example, in the system M/G/m/m we have constant mean delay $T = \bar{x}$ for all values of the input rate λ. In fact, the design variable λ is not equivalent to throughput in this case.

7.3 Combined Loss and Delay Systems

Let us augment the pure loss model analyzed in the section 7.2 by allowing (a finite number of) messages to wait in queue if all m channels are occupied. That is, we add a finite buffer and model the resulting system as an M/M/m/K queue (see Figure 7.3).

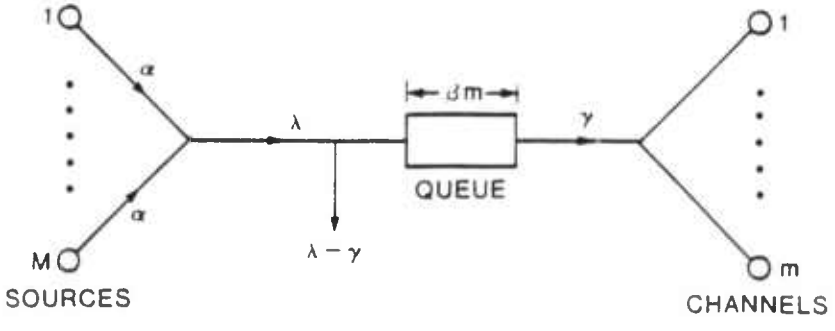

Figure 7.3 The M/M/m/K System

At most K messages may be in the system at one time, with at most $K - m$ of them in the queue. Any message that arrives to find all m channels busy and the buffer filled is lost (blocked). We use the same variable definitions as for the pure loss system, and also express the buffer size as a multiple of m (to maintain the same system characteristics as $m \to \infty$). Thus we set $K - m = \beta m$ or $K = \beta m + m$. Note that β is not necessarily an integer, although it is a rational number. As in the analysis of the pure loss system, we will study the various system parameters as $m \to \infty$ and note the effect of the smoothing principle on them.

Again we first need to determine the limiting behavior of the blocking probability B, because other system variables are described in terms of it. To this end, we define

$$S_m(A) = \frac{\sum_{j=0}^{m} A^j / j!}{A^m / m!} \tag{7.19}$$

Recall that in Theorem 7.1 we showed

$$S_m(A) = \int_0^\infty \left(1 + \frac{y}{A}\right)^m e^{-y}\, dy \tag{7.20}$$

and also

$$\lim_{m \to \infty} S_m(A) = \lim_{m \to \infty} S_m(m\delta) = \begin{cases} \infty & 0 \le \delta \le 1 \\ \dfrac{1}{1 - \dfrac{1}{\delta}} & 1 < \delta \end{cases} \tag{7.21}$$

Using equation (7.21), we will examine B for the combined loss and delay system as $m \to \infty$. To this end, we note for an M/M/m/K system with parameters λ and μ that

$$p_j = p_0 \left(\frac{\lambda}{\mu}\right)^j \frac{1}{j!} \qquad 1 \le j \le m$$

$$p_j = p_0 \left(\frac{\lambda}{\mu}\right)^j \frac{1}{m!\, m^{j-m}} \qquad m+1 \le j \le K$$

Solving for p_0 we find

$$p_0 = \left[\sum_{j=0}^{m} \left(\frac{\lambda}{\mu}\right)^j \frac{1}{j!} + \sum_{j=m+1}^{K} \left(\frac{\lambda}{\mu}\right)^j \frac{1}{m!\, m^{j-m}}\right]^{-1}$$

In our case, $A = \lambda \bar{x} = \lambda/\mu$, and so the blocking probability $B = p_K$ is given by

$$B = \frac{\dfrac{A^K}{m!\, m^{K-m}}}{\sum_{j=0}^{m} \dfrac{A^j}{j!} + \sum_{j=m+1}^{K} \dfrac{A^j}{m!\, m^{j-m}}} \tag{7.22}$$

Let us prove the following theorem which was stated in [Klei79], but for which he presented no proof.

Theorem 7.5

For the M/M/m/K system, as δ remains constant and $m \to \infty$ we have

$$\lim_{m \to \infty} B = \begin{cases} 0 & 0 \le \delta \le 1 \\ 1 - \dfrac{1}{\delta} & 1 \le \delta \end{cases} \tag{7.23}$$

This is the same result as for the pure loss system, which shows that the finite buffer has no effect on B in the limit.

To prove Theorem 7.5, we once more find it easier to consider $1/B$. Recalling that $A/m = \delta$ and $K - m = \beta m$, we have from equation (7.22)

$$\frac{1}{B} = \frac{\sum_{j=0}^{m} \frac{A^j}{j!} + \frac{A^m}{m!} \sum_{j=1}^{K-m} \frac{A^j}{m^j}}{\frac{A^m}{m!} \frac{A^{K-m}}{m^{K-m}}} = \frac{\sum_{j=0}^{m} \frac{A^j}{j!} + \frac{A^m}{m!} \sum_{j=1}^{\beta m} \delta^j}{\frac{A^m}{m!} \delta^{\beta m}}$$

and so

$$\frac{1}{B} = \frac{\sum_{j=0}^{m} \frac{A^j}{j!}}{\frac{A^m}{m!}} \left(\frac{1}{\delta}\right)^{\beta m} + \sum_{j=1}^{\beta m} \left(\frac{1}{\delta}\right)^{\beta m - j}$$

Therefore

$$\frac{1}{B} = S_m(A) \left(\frac{1}{\delta}\right)^{\beta m} + \sum_{j=0}^{\beta m - 1} \left(\frac{1}{\delta}\right)^j$$

For $1/\delta \ge 1$, since

$$\lim_{m \to \infty} \frac{1}{B} \ge \sum_{j=0}^{\infty} \left(\frac{1}{\delta}\right)^j$$

clearly

$$\lim_{m \to \infty} \frac{1}{B} = \infty$$

For $1/\delta < 1$, using equation (7.21) we have

$$\lim_{m \to \infty} \frac{1}{B} = \left(\frac{1}{1 - \frac{1}{\delta}}\right) \cdot 0 + \sum_{j=0}^{\infty} \left(\frac{1}{\delta}\right)^j = \frac{1}{1 - \frac{1}{\delta}}$$

Thus

$$\lim_{m\to\infty}\frac{1}{B}=\begin{cases}\infty & 0\le\delta\le 1\\ \dfrac{1}{1-\dfrac{1}{\delta}} & 1\le\delta\end{cases}$$

and we have Theorem 7.5

$$\lim_{m\to\infty}B=\begin{cases}0 & 0\le\delta\le 1\\ 1-\dfrac{1}{\delta} & 1\le\delta\end{cases}$$

[Note: the above limit is taken on the infinite sequence of those m such that βm is integer. Otherwise one would have a non-integer buffer size. For example, if $\beta=11/5$ say, then we must only consider those systems as $m\to\infty$ for which m is a multiple of 5; e.g. $m=5, 10, 15, 20, 25$, etc.]

As in the case of the pure loss system, since $\rho=\delta(1-B)$ we find

Theorem 7.6

For the M/M/m/K system, as δ remains constant and $m\to\infty$ we have

$$\lim_{m\to\infty}\rho=\begin{cases}\delta & 0\le\delta\le 1\\ 1 & 1\le\delta\end{cases}\tag{7.24}$$

Limiting expressions corresponding to Theorems 7.3 and 7.4 are more difficult to obtain than for the pure loss system, since we no longer have a simple expression for T. Thus we must calculate T and $\bar{N}/m$ explicitly. We choose to concentrate on $\bar{N}/m$. Using the expressions for p_k calculated above for the M/M/m/K system, we have

$$\bar{N}=\sum_{j=1}^{K}j\,p_j=\frac{\sum_{j=1}^{m}\left(\frac{\lambda}{\mu}\right)^j\frac{1}{(j-1)!}+\sum_{j=m+1}^{K}j\left(\frac{\lambda}{\mu}\right)^j\frac{1}{m!\,m^{j-m}}}{\sum_{j=0}^{m}\left(\frac{\lambda}{\mu}\right)^j\frac{1}{j!}+\sum_{j=m+1}^{K}\left(\frac{\lambda}{\mu}\right)^j\frac{1}{m!\,m^{j-m}}}$$

In our case, $A=\lambda\bar{x}=\lambda/\mu$, which yields the expression

$$\bar{N}=\frac{A\sum_{j=1}^{m}\frac{A^{j-1}}{(j-1)!}+\sum_{j=m+1}^{K}jA^j\frac{1}{m!\,m^{j-m}}}{\sum_{j=0}^{m}\frac{A^j}{j!}+\sum_{j=m+1}^{K}A^j\frac{1}{m!\,m^{j-m}}}$$

Dividing by m gives

$$\frac{\bar{N}}{m} = \frac{\frac{A}{m}\sum_{j=0}^{m-1}\frac{A^j}{j!} + \frac{1}{m}\sum_{j=m+1}^{K} jA^j\frac{1}{m!\,m^{j-m}}}{\sum_{j=0}^{m}\frac{A^j}{j!} + \sum_{j=m+1}^{K}A^j\frac{1}{m!\,m^{j-m}}}$$

or

$$\frac{\bar{N}}{m} = \frac{\frac{A}{m}\left(\sum_{j=0}^{m}\frac{A^j}{j!} - \frac{A^m}{m!}\right) + \sum_{j=1}^{K-m}\frac{j+m}{m}A^{j+m}\frac{1}{m!\,m^j}}{\sum_{j=0}^{m}\frac{A^j}{j!} + \sum_{j=1}^{K-m}A^{j+m}\frac{1}{m!\,m^j}}$$

Recalling that $K - m = \beta m$ and dividing numerator and denominator by $A^m/m!$ yields

$$\frac{\bar{N}}{m} = \frac{\frac{A}{m}\left[S_m(A) - 1\right] + \sum_{j=1}^{\beta m}\frac{j+m}{m}\left(\frac{A}{m}\right)^j}{S_m(A) + \sum_{j=1}^{\beta m}\left(\frac{A}{m}\right)^j}$$

Since $A/m = \delta$, we have

$$\frac{\bar{N}}{m} = \frac{\delta\left[S_m(A) - 1\right] + \sum_{j=1}^{\beta m}\delta^j + \frac{1}{m}\sum_{j=1}^{\beta m} j\delta^j}{S_m(A) + \sum_{j=1}^{\beta m}\delta^j} \tag{7.25}$$

We obtain another expression for $\bar{N}/m$ by dividing the numerator and denominator of equation (7.25) by $\delta^{\beta m}$, which gives

$$\frac{\bar{N}}{m} = \frac{\delta\left[S_m(A) - 1\right]\left(\frac{1}{\delta}\right)^{\beta m} + \sum_{j=1}^{\beta m}\left(\frac{1}{\delta}\right)^{\beta m - j} + \frac{1}{m}\sum_{j=1}^{\beta m} j\left(\frac{1}{\delta}\right)^{\beta m - j}}{S_m(A)\left(\frac{1}{\delta}\right)^{\beta m} + \sum_{j=1}^{\beta m}\left(\frac{1}{\delta}\right)^{\beta m - j}}$$

or

$$\frac{\bar{N}}{m} = \frac{\delta\left[S_m(A) - 1\right]\left(\frac{1}{\delta}\right)^{\beta m} + \sum_{j=0}^{\beta m-1}\left(\frac{1}{\delta}\right)^j + \frac{1}{m}\sum_{j=0}^{\beta m-1}(\beta m - j)\left(\frac{1}{\delta}\right)^j}{S_m(A)\left(\frac{1}{\delta}\right)^{\beta m} + \sum_{j=0}^{\beta m-1}\left(\frac{1}{\delta}\right)^j}$$

Therefore

$$\frac{\bar{N}}{m} = \frac{\delta\left[S_m(A) - 1\right]\left(\frac{1}{\delta}\right)^{\beta m} + (1+\beta)\sum_{j=0}^{\beta m-1}\left(\frac{1}{\delta}\right)^j - \frac{1}{m}\sum_{j=1}^{\beta m-1} j\left(\frac{1}{\delta}\right)^j}{S_m(A)\left(\frac{1}{\delta}\right)^{\beta m} + \sum_{j=0}^{\beta m-1}\left(\frac{1}{\delta}\right)^j} \tag{7.26}$$

We will use equations (7.25) and (7.26) (depending on the value of δ) to find the limiting value of $\bar{N}/m$, and then of $T/\bar{x}$ and power P. In the process we will correct three equations from [Klei79]; the cases $0 \leq \delta < 1$ and $1 < \delta$ were stated correctly there, but the case $\delta = 1$ was incorrectly stated. We now prove

Theorem 7.7

For the system M/M/m/K, as δ remains constant and $m \to \infty$ we have

$$\lim_{m \to \infty} \frac{\bar{N}}{m} = \begin{cases} \delta & 0 \leq \delta < 1 \\ 1 + \frac{\beta}{2} & 1 = \delta \\ 1 + \beta & 1 < \delta \end{cases} \tag{7.27}$$

First consider the case $\delta < 1$. From equation (7.25) we have

$$\frac{\bar{N}}{m} = \frac{\delta\left[1 - \frac{1}{S_m(A)}\right] + \frac{1}{S_m(A)} \sum_{j=1}^{\ell m} \delta^j + \frac{1}{m\, S_m(A)} \sum_{j=1}^{\ell m} j\, \delta^j}{1 + \frac{1}{S_m(A)} \sum_{j=1}^{\ell m} \delta^j}$$

Recalling that $\lim_{m \to \infty} S_m(A) = \infty$ for $\delta < 1$, we have

$$\lim_{m \to \infty} \frac{\bar{N}}{m} = \frac{\delta[1 - 0] + 0 \cdot \frac{\delta}{1 - \delta} + 0 \cdot \frac{\delta}{(1 - \delta)^2}}{1 + 0 \cdot \frac{\delta}{1 - \delta}}$$

or

$$\lim_{m \to \infty} \frac{\bar{N}}{m} = \delta$$

Now consider the case $\delta > 1$. In this case

$$\lim_{m \to \infty} \left(\frac{1}{\delta}\right)^{\ell m} = 0$$

and

$$\lim_{m \to \infty} S_m(A) = \frac{1}{1 - \frac{1}{\delta}} = \frac{\delta}{\delta - 1}$$

Hence equation (7.26) yields

$$\lim_{m\to\infty}\frac{\bar{N}}{m}=\frac{\delta\left[\frac{\delta}{\delta-1}-1\right]\cdot 0+(1+\beta)\cdot\frac{1}{1-(1/\delta)}-0\cdot\frac{1/\delta}{[1-(1/\delta)]^2}}{\left[\frac{\delta}{\delta-1}\right]\cdot 0+\frac{1}{1-(1/\delta)}}$$

or

$$\lim_{m\to\infty}\frac{\bar{N}}{m}=1+\beta$$

As might be expected, in this saturated situation the queue remains filled.

Finally, let us consider the numerically more difficult case of $\delta=1$. Again we use equation (7.25). Since $\delta=1$, we have $A=m$. Thus

$$\frac{\bar{N}}{m}=\frac{S_m(m)-1+\beta m+\frac{1}{m}\sum_{j=1}^{\beta m}j}{S_m(m)+\beta m}$$

or

$$\frac{\bar{N}}{m}=\frac{S_m(m)-1+\beta m+\frac{1}{m}\left[\frac{\beta m(1+\beta m)}{2}\right]}{S_m(m)+\beta m}$$

Therefore, we have

$$\frac{\bar{N}}{m}=\frac{S_m(m)-1+\beta m+\frac{\beta}{2}(1+\beta m)}{S_m(m)+\beta m}$$

or

$$\frac{\bar{N}}{m}=\frac{S_m(m)+\frac{\beta}{2}-1+\beta m\left(1+\frac{\beta}{2}\right)}{S_m(m)+\beta m}$$

This yields

$$\frac{\bar{N}}{m}=\frac{\frac{S_m(m)}{m}+\frac{\frac{\beta}{2}-1}{m}+\beta\left(1+\frac{\beta}{2}\right)}{\frac{S_m(m)}{m}+\beta}$$

We now claim that

$$\lim_{m\to\infty}\frac{S_m(m)}{m}=0 \qquad (7.28)$$

Assume this claim (it will be shown later). Then

$$\lim_{m\to\infty}\frac{\bar{N}}{m}=\lim_{m\to\infty}\frac{\dfrac{S_m(m)}{m}+\dfrac{\frac{\beta}{2}-1}{m}+\beta\left(1+\dfrac{\beta}{2}\right)}{\dfrac{S_m(m)}{m}+\beta}$$

or

$$\lim_{m\to\infty}\frac{\bar{N}}{m}=\frac{0+0+\beta\left(1+\frac{\beta}{2}\right)}{0+\beta}$$

We finally obtain

$$\lim_{m\to\infty}\frac{\bar{N}}{m}=1+\frac{\beta}{2}$$

Thus we note the interesting result that, on the average, as $m\to\infty$ for $\delta=1$ the queue remains half-full.

To complete the proof we need only verify equation (7.28), namely,

$$\lim_{m\to\infty}\frac{S_m(m)}{m}=0$$

To this end, using equation (7.20), we first rewrite $S_m(m)/m$ as

$$\frac{S_m(m)}{m}=\int_0^\infty\frac{\left(1+\frac{y}{m}\right)^m e^{-y}}{m}\,dy=\int_0^\infty\frac{\left[\left(1+\frac{y}{m}\right)e^{-\frac{y}{m}}\right]^m}{m}\,dy$$

Now make the change of variable $x=y/m$, so that $dx=dy/m$. This yields

$$\frac{S_m(m)}{m}=\int_0^\infty\left[(1+x)\,e^{-x}\right]^m dx$$

Define the functions (for $m=1,2,\ldots$)

$$f_m(x)\triangleq\left[(1+x)\,e^{-x}\right]^m$$

Note that

$$f_1(x)=(1+x)\,e^{-x}\begin{cases}=1 & x=0\\ <1 & x>0\end{cases}$$

Therefore

$$\left[(1+x)\,e^{-x}\right]^m\geq\left[(1+x)\,e^{-x}\right]^{m+1}$$

which means that

$$f_m(x) \geq f_{m+1}(x)$$

Let $f(x) \triangleq \lim_{m \to \infty} f_m(x)$. Then

$$f(x) = \begin{cases} 1 & x = 0 \\ 0 & x > 0 \end{cases}$$

and so

$$\int_0^\infty f(x)\, dx = 0$$

Now since $f_1 \geq f_2 \geq f_3 \geq \ldots$ and

$$\int_0^\infty f_1(x)\, dx = \int_0^\infty (1+x)\, e^{-x}\, dx = 2 < \infty$$

the functions f_m are dominated by the integrable function f_1. Hence the Lebesgue Dominated Convergence Theorem may be applied to interchange limit and integral and give

$$\lim_{m \to \infty} \int_0^\infty f_m(x)\, dx = \int_0^\infty \lim_{m \to \infty} f_m(x)\, dx = \int_0^\infty f(x)\, dx = 0$$

Thus

$$\lim_{m \to \infty} \frac{S_m(m)}{m} = \lim_{m \to \infty} \int_0^\infty f_m(x)\, dx = 0$$

and the claim (equation (7.28)) is proved.

We now can find the limiting values of T and the power function P. Since $\bar{N}/m = \rho(T/\bar{x})$, Theorems 7.6 and 7.7 yield

Theorem 7.8

For the system M/M/m/K, as δ remains constant and $m \to \infty$ we have

$$\lim_{m \to \infty} \frac{T}{\bar{x}} = \begin{cases} 1 & 0 \leq \delta < 1 \\ 1 + \dfrac{\beta}{2} & 1 = \delta \\ 1 + \beta & 1 < \delta \end{cases} \tag{7.29}$$

Next, using the definition of power that includes blocking given in equation (7.17), we find from Theorems 7.5, 7.6, and 7.8

Theorem 7.9

For the system $M/M/m/K$, as δ remains constant and $m \to \infty$ we have

$$\lim_{m \to \infty} P = \begin{cases} \delta & 0 \leq \delta < 1 \\ \dfrac{1}{1+\dfrac{\beta}{2}} & 1 = \delta \\ \dfrac{1}{(1+\beta)\,\delta} & 1 < \delta \end{cases} \tag{7.30}$$

This last expression gives us an interesting example of a (discontinuous) power function which has no maximum point. The point at which we might think the maximum would occur ($\delta = 1$) is a point of discontinuity of P.

7.4 An Application to a Packetized Voice Network

Recent advances in communications technology (e.g. fiberoptics) promise the capability of transmitting messages across networks at very great speeds (on the order of gigabits per second). A paper by Roberts [Robe82] suggests that much of this huge bandwidth may be wasted if we restrict the use of such networks to conventional data sources (facsimile, electronic mail, etc.). One possible area of application for these new advances is packetized voice. In this section we study a simple model of speech where we specifically do not allow packets to queue. This model is appropriate since random blocking of voice packets is claimed not to seriously affect speech. Instead one usually wishes minimal variance in delay and a constant rate of packets through the net. Speech is capable of sustaining some loss of packets; however, it will degrade appreciably if large numbers of consecutive packets are blocked.

The environment we consider consists of many users utilizing the capacity of a communications channel whose bandwidth has been split into many individual subchannels. We model this situation as an M/G/m/m queueing system, and thus we may utilize the results of section 7.2 above. Therefore we have a pure loss system consisting of m channels (servers) with no queueing of incoming packets. Any packet which arrives to find all m channels busy will be lost (or blocked). As in section 7.2, we assume that arrivals occur from M users, each sending packets at a Poisson rate of α. Thus the total arrival rate of speech packets is Poisson at a rate λ where $\lambda = M\alpha$. The service time distribution is assumed to be general with mean $\bar{x}$. Here $\bar{x} = \bar{b}/C$ where $\bar{b}$ is the mean packet length in bits and C is the channel capacity in bits per second. The total applied load to the system, A, satisfies $A = \lambda\bar{x} = M\alpha\bar{x}$. The offered load per channel δ is $\delta = A/m = (\lambda\bar{x})/m = (M/m)\alpha\bar{x}$. Performance measures of interest include the probability that a packet is lost or blocked. Thus the parameters of interest are given in terms of the blocking probability $B = B_m(A)$. We see that the actual arrival rate to the system is $\gamma = \lambda(1-B)$, and thus $\gamma\bar{x} = \lambda(1-B)\bar{x} = A(1-B)$ is the total carried load. The efficiency ρ of each channel is $\rho = (\gamma\bar{x})/m = \delta(1-B)$.

We now extend the above model to a series network of such M/G/m/m nodes. That is, we consider a tandem of N channels, each split into m subchannels. The output of an M/G/m/m system is Poisson if arrival points of blocked customers are considered as departures [Cohe76]. However, departures of non-blocked customers (those that have been successfully served) are not Poisson. As a rough approximation we assume, for simplicity, both that the blocked traffic has been exactly balanced by introducing additional traffic at each successive node at a rate of $\lambda - \gamma$, and that the resulting combined traffic (at rate λ) is Poisson. If we examine a large network of such nodes, we may consider our N-node tandem as a path (virtual channel) through the packetized network.

Define $P_S^{(N)}$ as the probability that a packet is not blocked along the N-node tandem. Using the above approximation, we have that

$$P_S^{(N)} = [P_S]^N$$

where, recall, $P_S = 1 - B$ is the success probability for a single M/G/m/m system (studied in section 7.2). The behavior of $P_S^{(N)}$ for $N = 5$ hops is plotted in Figure 7.4. For m large we see a dramatic deterministic behavior.

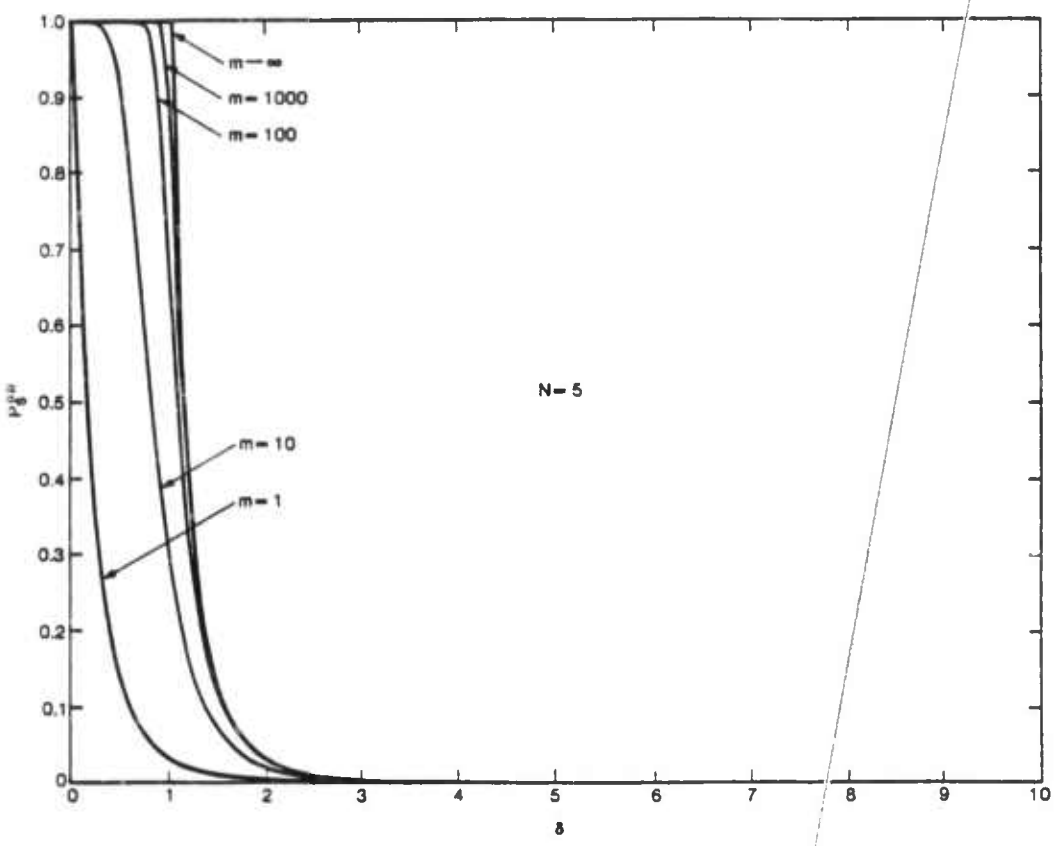

Figure 7.4 Success Probability for a Path with 5 Hops

Using the result for a single node from equation (7.14), we may find the limiting probability that a packet will successfully traverse a path of N hops as

$$\lim_{m\to\infty} P_S^{(N)} = \begin{cases} 1 & 0 \le \delta \le 1 \\ \dfrac{1}{\delta^N} & 1 \le \delta \end{cases} \tag{7.31}$$

We therefore have, for very long tandems, that the limit is

$$\lim_{N\to\infty}\lim_{m\to\infty} P_S^{(N)} = \begin{cases} 1 & 0 \le \delta \le 1 \\ 0 & 1 < \delta \end{cases} \tag{7.32}$$

For an infinite tandem ($N \to \infty$), there is a critical value of offered load per channel $\delta_c = 1$ such that a packet will either eventually surely be blocked or will surely be successful, depending on whether the system is overloaded ($\delta > \delta_c$) or not ($\delta \le \delta_c$) respectively. In Figure 7.5 the limiting case of $m \to \infty$ is plotted for $N = 1, 5, 10, 15, 20, \infty$. The 0-1 behavior of the success probability is quite evident as N becomes large.

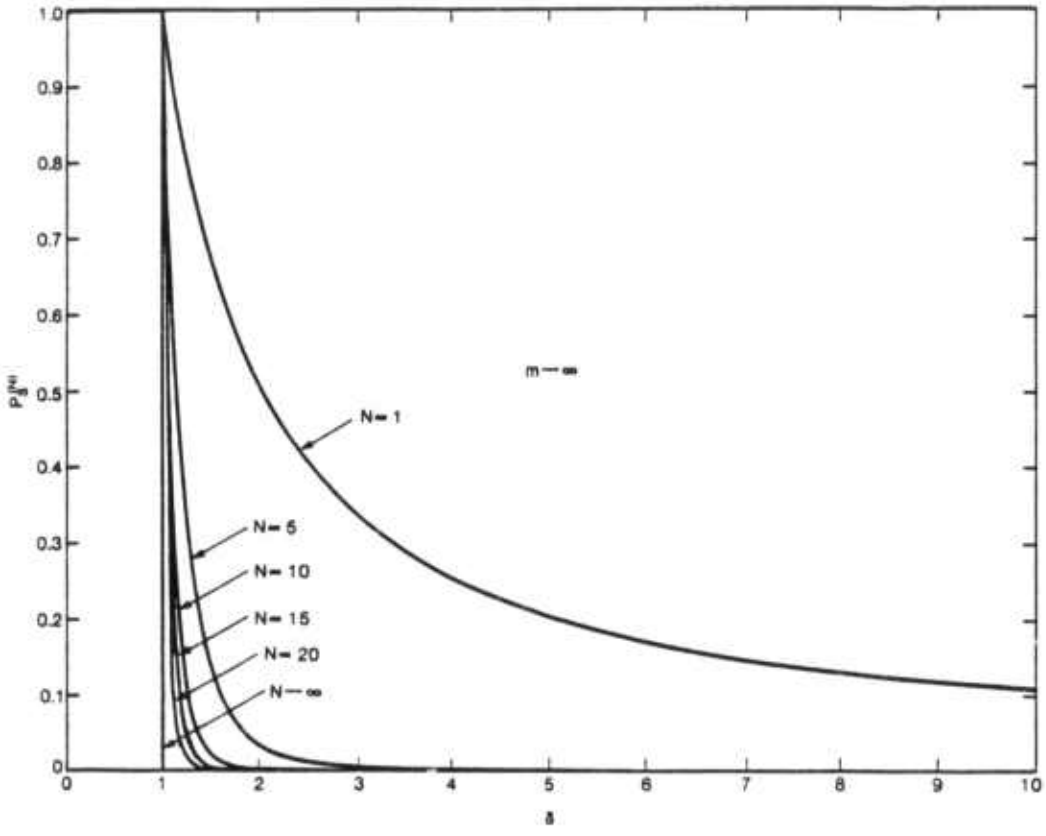

Figure 7.5 Success Probability for the Limiting Case of m

We next consider the number of hops until a packet is blocked (i.e., dropped). Let $\tilde{n}$ be a random variable representing the number of hops a packet successfully traverses along the path, and let B be the blocking probability for a single M/G/m/m system. Thus, assuming N nodes in the chain, $\tilde{n} = k$ with probability $(1-B)^k B$ for $0 \le k \le N-1$ and $\tilde{n} = N$ with probability $(1-B)^N$ $(= P_S^{(N)})$. For $B = 0$ we have $P_S^{(N)} = 1$, and a packet will surely traverse all N hops on the path (B can be zero, at least in the limiting case of m, by Theorem 7.1).

Now assume $B > 0$. In this case $\bar{n}$, the average number of hops a packet successfully traverses, is given by

$$\bar{n} = \sum_{k=1}^{N-1} k(1-B)^k B + N(1-B)^N = (1-B)B\sum_{k=1}^{N-1} k(1-B)^{k-1} + N(1-B)^N$$

We note that

$$\sum_{k=1}^{N-1} kw^{k-1} = \frac{d}{dw}\sum_{k=0}^{N-1} w^k = \frac{d}{dw}\frac{1-w^N}{1-w} = \frac{1-Nw^{N-1}+(N-1)w^N}{(1-w)^2}$$

for any w. Setting $w = 1-B$, we have

$$\bar{n} = (1-B)B\,\frac{1-N(1-B)^{N-1}+(N-1)(1-B)^N}{[1-(1-B)]^2} + N(1-B)^N$$

Therefore

$$\bar{n} = \frac{(1-B)[1-(1-B)^N]}{B}$$

For $N \to \infty$ (and $B > 0$) we obtain the simple expression

$$\bar{n}_\infty \triangleq \lim_{N\to\infty} \bar{n} = \sum_{k=1}^{\infty} k(1-B)^k B = \frac{1-B}{B} \tag{7.33}$$

Using this latter equation, we have $\bar{n}_\infty = 1/\delta$ when $m = 1$, since $B = \delta/(1+\delta)$ for M/G/1/1. The limiting case of $m \to \infty$ yields (using Theorem 7.1)

$$\bar{n}_\infty = \begin{cases} \infty & 0 \le \delta \le 1 \text{ (never blocked)} \\ \dfrac{1}{\delta - 1} & 1 < \delta \end{cases} \tag{7.34}$$

Again, for $m \to \infty$, we observe the shifted behavior of an important system parameter from the $m = 1$ case when $1 < \delta$. Also note that as $\delta \to \infty$ then $\bar{n}_\infty \to 0$, and a packet is blocked at the first hop. For $0 \le \delta \le 1$ a packet is never blocked and successfully traverses the entire path. Figure 7.6 gives a plot of $\bar{n}_\infty$ versus δ for various values of m. We see that the curve for $m = 100$ gives values very near the limiting case of $m \to \infty$.

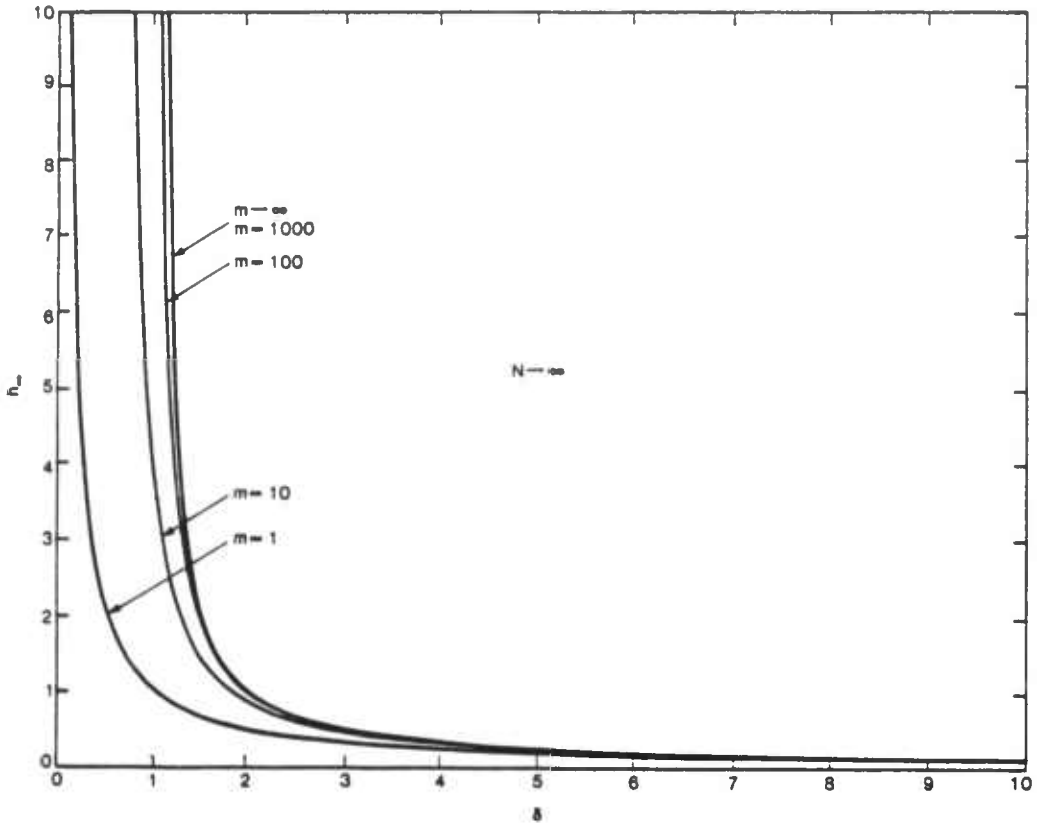

Figure 7.6 Mean Number of Successful Hops for the Limiting Case of N

Examining equation (7.34) in greater detail, we observe that if there is a 1/100 overload on the system ($\delta = 1 + 1/100$) then $\bar{n}_\infty = 100$, while if there is a 1/10 overload then $\bar{n}_\infty = 10$. This provides us with a nice rule of thumb relating the fraction of overload with the average number of successful hops in large nets. We now consider a tandem of N nodes each modeled as an M/G/m/m system with $m \to \infty$. Suppose the system is in an overloaded state with the offered load equal to $\delta = 1 + 1/K$. Thus, from equation (7.34), this value of offered load to the system would yield an average of K successful hops per packet in an infinite tandem. That is, $\bar{n}_\infty = K$ for $\delta = 1 + 1/K$. Since we are examining the case when $m \to \infty$, we see by equation (7.14) that the probability of success for a single hop is $P_S = 1/\delta$. The probability of success in the N hop system is therefore

$$P_S^{(N)} = [P_S]^N = \left(\frac{1}{\delta}\right)^N$$

or

$$P_S^{(N)} = \frac{1}{\left(1 + \frac{1}{K}\right)^N}$$

For $K = 1$, then $\delta = 2$, and thus $P_S^{(N)} = 1/2^N$. For K large, we first write $N = \alpha K$ (so that $\alpha = N/K$). Then we have

$$P_S^{(N)} = \frac{1}{\left(1 + \frac{1}{K}\right)^N} = \frac{1}{\left(1 + \frac{\alpha}{N}\right)^N}$$

and so for δ near 1 we find

$$P_S^{(N)} \to \frac{1}{e^{\alpha}} = e^{-\alpha}$$

This gives an estimate of the fraction of successful traffic in an overloaded N hop network from the parameters of the limiting system.

In this section we have introduced a model of a packetized voice network with many users and many voice channels which incorporates blocking. The model utilized the rough approximation of balancing blocked traffic for each node of the network with input traffic from the other nodes so that the Poisson assumption could be used. Our primary interest involved studying the interplay of input parameters λ (input rate), $\bar{x}$ (mean service time), m (number of channels), and N (number of nodes) in such a model. The limiting behavior of a single node in such a pure loss network which was analyzed in section 7.2 was extended to a path through the network. We obtained intuition into the deterministic behavior of large nets (which occurs as a consequence of the smoothing effect of the law of large numbers). The various plots in the paper also graphically illustrate this behavior. We see that there are mixtures of the input parameters of interest (λ, $\bar{x}$, m, and N) which give excellent behavioral characteristics. In the extreme limiting case of $m \to \infty$ and $N \to \infty$ we find a critical threshold $\delta_c = 1$ of offered load per channel such that for $\delta \leq \delta_c$ the probability of success is one and all packets are successful, while for $\delta > \delta_c$ the success probability is zero and all packets are lost. The limiting case of $m \to \infty$ and $N \to \infty$ was also used to estimate the fraction of traffic which will successfully traverse a path of K hops in an overloaded, large, pure loss network.

This concludes our analysis of multiple server systems, with and without blocking. In general, it presented more difficulties than for single server systems. But for very large systems (i.e., large numbers of servers), the smoothing effect of the law of large numbers takes hold and deterministic behavior results. Exploiting this behavior, a power function first introduced by Kleinrock which includes blocking was studied for large nets. It was found to peak at the right value of offered load for the limiting pure loss system.

CHAPTER 8
Conclusions and Suggestions for Further Work

As previously discussed, this research addresses the problem of "where should one operate a computer network". In an attempt to partially answer the above (rather vague) question in this dissertation, we have been led to the consideration of throughput-delay tradeoff functions. One possible approach which was taken in this work involves the optimization of the various notions of power and generalized power defined above. We have seen for simple computer networks (those leading to simple optimization problems), that the choice of the operating point which maximizes power leads to several intuitively pleasing results. Rules of thumb such as "keep the pipe full", and the invariances shown by $\overline{N}^*$, the average number at optimum power, are examples. However, optimizing power for more complex network problems was difficult and also led to unwanted side effects. Applying Kleinrock's definition of power (instead of that of Giessler or Nakamura) to these complicated network problems did eliminate some of the undesirable properties of the optimal power point.

In applying these results to real networks, one use could be to provide a network operating point which would be the target of a flow control procedure. The parameter $\overline{N}^*$ could be used as a handle that window flow control system design would deal with. Thus the results about $\overline{N}^*$ obtained in this research are implementable in real systems.

We now indicate future areas of research which are related to the work done in this dissertation. The analysis of power itself may continue in several directions. More complex optimization problems (such as PF1 and PF2 of chapter 4) using the conventional computer network model may be analyzed. It appears that the analysis becomes harder and difficulties in optimization may arise. A definition of power (in the spirit of that introduced by Kleinrock) is needed which has physical meaning as a throughput-delay tradeoff function and whose "nice" properties extend beyond the simple network optimization problems to the multi-variable case.

We may also extend the analysis of power to models of computer phenomena other than the traditional wire networks studied above. A step in this direction was the study of blocking models in chapter 7. One useful set of models are those of broadcast networks, be they packet radio, satellite, or broadcast cable. Local networks are becoming particularly important. The delay characteristics of the various broadcast protocols (ALOHA, CSMA, etc.) are usually rather complicated. A method of analysis introduced by Kleinrock [Klei77] provides an approximation to delay (the ZAP approximation) which could be studied from the power perspective.

Another set of models to investigate in terms of power are those of time-shared systems. These models lead naturally to the study of power for queueing disciplines other than first-come-first-served. For example, many time-sharing models assume a processor-sharing discipline [Lave83]. The classic work of Baskett et al [Bask75] on networks of queues which allow various types of nodes other than first-come-first-served may be useful. A recent paper by Cohen [Cohe79] introduces a discipline he calls *generalized processor sharing,* which could also be analyzed from the power point of view. The M/G/m/m system studied in chapter 7 is, in fact, a special example of a generalized processor sharing system.

Computer systems performance modeling from the point of view of power not only leads to the consideration of other queueing disciplines, but also to the analysis of power for closed networks of queues. The networks considered in this work were all open networks, but many computer phenomena are modeled using closed (and mixed) networks [Lave83]. Even in the case of certain conventional wire networks, the model employed consists of closed chains. An example is the model developed by Reiser [Reis79] for a virtual circuit network with window flow control. Some preliminary work on power for closed networks has already appeared in the literature [Bhar80, Hsie83], but the networks considered are topologically (and mathematically) simple.

In addition to these network models for analyzing computer performance, other models for certain computer phenomena have also been introduced. One particular area of recent interest is in the modeling of distributed systems. With the rapid evolution of distributed processing technology, models of distributed systems have become increasingly important. One technique of modeling these systems, stochastic Petri nets, takes into account throughput and delay measures and was first introduced by Molloy [Moll81]. These stochastic Petri net models could perhaps be analyzed with respect to power.

Another area of research (in the spirit of chapter 4) concerns performance criteria other than the power functions studied above. Three types of power functions (and their extensions to generalized power) were analyzed in this dissertation, but each definition was not without its drawbacks (either in the physical meaning of the function or in the difficulty of optimizing it in certain cases). Other performance measures which yield a throughput delay tradeoff may be defined and studied. An example is given by Kermani and Kleinrock in [Kerm80]. There they introduce a function which is simply a linear combination of throughput and delay ($\alpha\gamma - T$, $\alpha \geq 0$). We note that a preference for throughput (or delay) can be indicated by simply adjusting the parameter α, similar to the generalized power family. Another example appears in [Yosh81] where they extend their single node power function P_N to a general network topology. However, their extension is based on the channel flows $\{\lambda_i\}$ and thus the total internal network traffic, whereas all the power functions considered in this dissertation are based on the throughput and thus the traffic matrix $\{\gamma_{jk}\}$. Therefore, their measure does not have physical meaning as a throughput-delay tradeoff function, and it consequently does not seem as useful as the usual power definitions.

Another approach (also in the spirit of chapter 4) is to change the constraints and/or the decision variables of the problem under study. An example appears in [Tann81] where the power of individual users is maximized under the constraint that all users have the same power. This does not seem very interesting, however, since all users are constrained by the maximum power of the slowest user.

In discussing performance measures related to the original notion of power it may be natural to ask whether these measures can in any way be made to correspond to the concept of power found in electrical circuit theory. From our results we can see that, for an M/M/1 tandem, the power of the system P is related to the power P_i of the individual nodes as

$$\frac{1}{P} = \frac{1}{P_1} + \cdots + \frac{1}{P_M}.$$

This looks suspiciously like a result for the admittance of a parallel electrical circuit. In fact the measure of power presented in [Yosh81] was obtained by noting a correspondence between an original computer network and an electrical circuit network. Thus the wealth of techniques from electrical circuit theory (e.g. Kirchoff's Laws) may perhaps be useful in analyzing computer networks.

We now address a most important direction of future research, namely the implementation in real systems of the "appropriate" operating point. In the optimization problems studied in this dissertation, we assume a fixed environment with a centralized ability to achieve an optimal operating point. But in practice, dynamic behavior may occur and a distributed algorithm may be needed. Such an algorithm was given in [Gall77] when the performance measure was delay, but we know of no such algorithm for optimizing global power. In fact, as we noted in chapter 3, Jaffe [Jaff81] gave an example of a network in which local power (where all nodes of the network use only local information in optimizing power) and global power differ. That is, the resulting local power value is not globally optimal. Thus globally maximizing power using a "decentralizable" algorithm may not be possible for certain networks.

Since local power and global power do not, in general, give identical results, another set of questions involve the *approximation* of global power using the intuition and rules of thumb derived from the local point of view. Flow control strategies that use only local information (in the calculation of window size, for example) may be used to somehow approximate a good global operating point. A first step in this direction was given by Bharath-Kumar and Jaffe in [Bhar81]. They developed a heuristic "greedy" distributed algorithm where, one by one, each user maximizes his own power based upon current information. As pointed out by Yemini [Yemi81], this type of *selfish* optimization is related to the well-known approach called Pareto optimality. Yemini examines the optimization of power from this point of view, and it appears that the theory of Pareto optimality may be useful in the analysis of distributed algorithms for optimizing power (or any other favorite performance measure).

Another important performance criterion which has appeared in this work is fairness. Various concepts of fairness have been defined [Gerl82, Wong82] besides the one used in this dissertation. One may try to extend the definition of power (the product of powers introduced by Bharath-Kumar and Jaffe [Bhar81] is an example) or change the constraints of the problem to yield fair solutions.

In this dissertation, we *analyzed* networks from the power point of view always assuming that the topology and the channel capacities were known. Instead, we may consider the capacities $\{C_i\}$ and/or the network topology to be decision variables, thus creating new power problem formulations (recall that we proceeded slightly in that direction in chapter 4 when we noted that the usual capacity assignment problem did not change when power was used as the objective function rather than delay). Such extended problem formulations naturally lead to another possible area of research, namely the *design* of computer networks with power as a performance measure. Thus we can assume that the channel capacities and the topology are variables and try to find networks with a structure which optimizes power.

We end by emphasizing the difficulty of the question asked at the beginning of this dissertation, namely, finding an appropriate operating point for a computer network. Hopefully this research has contributed toward some clarification of this question.

References

[Bask75] F. Baskett, K. M. Chandy, R. R. Muntz, and F. Palacios-Gomez, "Open, Closed and Mixed Networks of Queues with Different Classes of Customers," *Journal of the ACM* **22**(2), pp.260-284 (April 1975).

[Bhar80] K. Bharath-Kumar, "Optimum End-to-End Flow Control in Networks," *Conference Record, International Conference on Communications*, pp.23.3.1-23.3.6 (June 1980).

[Bhar81] K. Bharath-Kumar and J. M. Jaffe, "A New Approach to Performance Oriented Flow Control," *IEEE Transactions on Communications* **COM-29**(4), pp.427-435 (April 1981).

[Cant74] D. G. Cantor and M. Gerla, "Optimal Routing in a Packet-Switched Computer Network," *IEEE Transactions on Computers* **C-23**(10), pp.1062-1069 (October 1974).

[Cohe69] J. W. Cohen, *The Single Server Queue*, Wiley-Interscience, New York (1969).

[Cohe76] J. W. Cohen, *On Regenerative Processes in Queueing Theory*, Springer-Verlag, Berlin (1976).

[Cohe79] J. W. Cohen, "The Multiple Phase Service Network with Generalized Processor Sharing," *Acta Informatica* **12**(3), pp.245-284 (1979).

[Cour80] P. J. Courtois and P. Semal, "A Flow Assignment Algorithm Based on the Flow Deviation Method," *Proceedings of the Fifth International Conference on Computer Communication*, pp.77-83 (October 1980).

[Frat73] L. Fratta, M. Gerla, and L. Kleinrock, "The Flow Deviation Method: An Approach to Store-and-Forward Communication Network Design," *Networks* **3**, pp.97-133 (1973).

[Gall77] R. G. Gallager, "A Minimum Delay Routing Algorithm Using Distributed Computation," *IEEE Transactions on Communications* **COM-25**(1), pp.73-85 (January 1977).

[Geof70] A. M. Geoffrion, "Elements of Large-Scale Mathematical Programming," *Management Science* **16**(11), pp.652-691 (July 1970).

[Gerl77] M. Gerla and L. Kleinrock, "On the Topological Design of Distributed Computer Networks," *IEEE Transactions on Communications* **COM-25**(1), pp.48-60 (January 1977).

[Gerl80] M. Gerla and L. Kleinrock, "Flow Control: A Comparative Survey," *IEEE Transactions on Communications* **COM-28**(4), pp.553-574 (April 1980).

[Gerl80a] M. Gerla and P. O. Nilsson, "Routing and Flow Control Interplay in Computer Networks," *Proceedings of the Fifth International Conference on Computer Communication*, pp.84-89 (October 1980).

[Gerl82] M. Gerla and M. Staskauskas, "Fairness in Flow Controlled Networks," *Journal of Telecommunication Networks* **1**(1), pp.29-38 (Spring 1982).

[Gies78] A. Giessler, J. Hanle, A. Konig, and E. Pade, "Free Buffer Allocation — An Investigation by Simulation," *Computer Networks* **1**(3), pp.191-204 (July 1978).

[Gran79] J. Grangé and M. Gien (eds.), *Flow Control in Computer Networks*, IFIP, North-Holland Publishing Company, Amsterdam (1979).

[Hsie83] W. Hsieh and B. Kraimeche, "Performance Analysis of an End-to-End Flow Control Mechanism in a Packet-Switched Network," *Journal of Telecommunication Networks* **2**(1), pp.103-116 (Spring 1983).

[Jack63] J. R. Jackson, "Jobshop-Like Queueing Systems," *Management Science* **10**(1), pp.518-521 (1963).

[Jaff81] J. M. Jaffe, "Flow Control Power is Non-Decentralizable," *IEEE Transactions on Communications* **COM-29**(9), pp.1301-1306 (September 1981).

[Kenn80] J. L. Kennington and R. V. Helgason, *Algorithms for Network Programming*, Wiley-Interscience, New York (1980).

[Kerm80] P. Kermani and L. Kleinrock, "Dynamic Flow Control in Store-and-Forward Computer Networks," *IEEE Transactions on Communications* **COM-28**(2), pp.263-271 (February 1980).

[Klei64] L. Kleinrock, *Communication Nets: Stochastic Message Flow and Delay*, McGraw-Hill, New York (1964). Out of print. Reprinted by Dover Publications, 1972.

[Klei75] L. Kleinrock, *Queueing Systems, Vol I., Theory*, Wiley-Interscience, New York (1975).

[Klei76] L. Kleinrock, *Queueing Systems, Vol II., Computer Applications*, Wiley-Interscience, New York (1976).

[Klei77] L. Kleinrock, "Performance of Distributed Multi-Access Computer-Communication Systems," *Information Processing 77, Proceedings of IFIP Congress 77*, pp.547-552 (August 1977).

[Klei78] L. Kleinrock and Y. Yemini, "An Optimal Adaptive Scheme for Multiple Access Broadcast Communication," *Conference Record, International Conference on Communications*, pp.7.2.1-7.2.5 (June 1978).

[Klei78a] L. Kleinrock, "On Flow Control in Computer Networks," *Conference Record, International Conference on Communications* **2**, pp.27.2.1-27.2.5 (June 1978).

[Klei79] L. Kleinrock, "Power and Deterministic Rules of Thumb for Probabilistic Problems in Computer Communications," *Conference Record, International Conference on Communications*, pp.43.1.1-43.1.10 (June 1979).

[Klei80] L. Kleinrock and P. Kermani, "Static Flow Control in Store-and-Forward Computer Networks," *IEEE Transactions on Communications* **COM-28**(2), pp.271-279 (February 1980).

[Lave83] S. S. Lavenberg (ed.), *Computer Performance Modeling Handbook*, Academic Press, New York (1983).

[Litt61] J. Little, "A Proof for the Queuing Formula: $L = \lambda W$," *Operations Research* **9**(3), pp.383-387 (May 1961).

[Luen73] D. G. Luenberger, *Introduction to Linear and Nonlinear Programming*, Addison-Wesley, Reading, Mass. (1973).

[Mang69] O. L. Mangasarian, *Nonlinear Programming*, McGraw-Hill, New York (1969).

[Marl78] W. H. Marlow, *Mathematics for Operations Research*, Wiley-Interscience, New York (1978).

[Mitt81] K. K. Mittal and A. N. Venetsanopoulos, "On the Dynamic Control of the Urn Scheme for Multiple Access Broadcast Communication Systems," *IEEE Transactions on Communications* **COM-29**(7), pp.962-970 (July 1981).

[Moll82] M. L. Molle, "On the Capacity of Infinite Population Multiple Access Protocols," *IEEE Transactions on Information Theory* **IT-28**(3), pp.396-401 (May 1982).

[Moll81] M. Molloy, "On the Integration of Delay and Throughput Measures in Distributed Processing Models," CSD Report No. 810921 (UCLA-ENG-8127), Computer Science Department, University of California, Los Angeles (September 1981). Ph.D. Dissertation.

[Nels82] R. Nelson, "Channel Access Protocols for Multi-Hop Broadcast Packet Radio Networks," CSD Report No. 820731 (UCLA-ENG-82-59), Computer Science Department, University of California, Los Angeles (July 1982). Ph.D. Dissertation.

[Reis79] M. Reiser, "A Queueing Network Analysis of Computer Communication Networks with Window Flow Control," *IEEE Transactions on Communications* **COM-27**(8), pp.1199-1209 (August 1979).

[Robe82] L. G. Roberts, "Packet Switching Economics," *Journal of Telecommunication Networks* **1**(3), pp.213-218 (Fall 1982).

[Rubi74] I. Rubin, "Communication Networks: Message Path Delays," *IEEE Transactions on Information Theory* **IT-20**(6), pp.738-745 (November 1974).

[Saue81] C. H. Sauer and K. M. Chandy, *Computer Systems Performance Modeling*, Prentice-Hall, Inc., Englewood Cliffs, New Jersey (1981).

[Schw80] M. Schwartz and T. E. Stern, "Routing Techniques Used in Computer Communication Networks," *IEEE Transactions on Communications* **COM-28**(4), pp.539-552 (April 1980).

[Sysk60] R. Syski, *Introduction to Congestion Theory in Telephone Systems*, Oliver and Boyd, London (1960).

[Tann81] K. Tanno, H. Kuniyoshi, T. Nakamura, and R. Sato, "A Flow Control Analysis Based on Measure of Power in Packet Switching Networks," *Conference Record, National Telecommunications Conference*, pp.E5.7.1-E5.7.6 (November 1981).

[Wong82] J. W. Wong, J. P. Sauvé, and J. A. Field, "A Study of Fairness in Packet-Switching Networks," *IEEE Transactions on Communications* **COM-30**(2), pp.346-353 (February 1982).

[Yemi81] Y. Yemini, "Selfish Optimization in Computer Networks," *Proceedings of the 20th IEEE Conference on Decision and Control*, pp.374-379 (December 1981).

[Yosh77] Y. Yoshioka, T. Nakamura, and R. Sato, "An Optimum Solution of the Queuing System," *Electronics and Communications in Japan* **60-B**(8), pp.590-591 (August 1977).

[Yosh81] Y. Yoshioka, T. Nakamura, and Y. Shigei, *Optimum Traffic in the Queueing System*, Department of Information Science, Tohoku University, Sendai, Japan (1981). Submitted to *IEEE Transactions on Communications*.